DOCUMENT RESUME

ED 175 945 TM 009 703

AUTHOR Murray, Steve; And Others
TITLE Title I Technical Issues as Threats to Internal Validity of Experimental and Quasiexperimental Designs.
PUB DATE Apr 79
NOTE 131p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979)

EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS Achievement Gains; Compensatory Education Programs; *Evaluation Methods; Models; Norm Referenced Tests; *Program Evaluation; *Research Design; Research Methodology; *Research Problems; Scoring; Testing Problems; Test Validity; *Validity
IDENTIFIERS Elementary Secondary Education Act Title I; Normal Curve Equivalent Scores; *RMC Models; *Title I Evaluation and Reporting System

ABSTRACT

Threats to the validity of the Title I Evaluation and Reporting System are covered in two parts: an annotated bibliography of reports concerned with technical issues, and a discussion of threats to validity--from the reporting system in general and from each model in particular. Threats common to all three evaluation models are reported: instrumentation effects involving scaling as related to internal validity; statistical power; multiple comparisons and the error rate problem; test reliability as related to statistical conclusion validity; aptitude--treatment interactions; inadequate preoperational explication of the treatment construct as related to construct validity of cause; and inadequate preoperational definition of effect construct and mono-operational bias as related to construct validity of effects. Specific problems associated with Model A, Model B, and Model C are reviewed, and suggestions are made for overcoming these problems. (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED175945

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Title I Technical Issues as Threats to Internal Validity of Experimental and Quasiexperimental Designs

Steve Murray Judy Arter

Bonnie Faddis

Northwest Regional Educational Laboratory

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Judy Arter

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

April 1979

Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 8-12, 1979. Session 6.12.

TM009 703

NWREL TAC

3/79
3260A

Introduction

The Title I Evaluation and Reporting System is the first nationwide attempt to standardize evaluation practice and obtain comparable evaluation results across programs. The system focuses on measuring the impact of Title I projects on academic achievement in mathematics, language arts and reading. The measure of impact, the treatment effect, is generated by applying one of three evaluation models according to its implementation guidelines. Comparable reporting is facilitated through a new standard score metric, the Normal Curve Equivalent (NCE). It is expected that programs receiving Title I funds will be required to use the Title I Evaluation and Reporting System in order to demonstrate the effects of their instruction.

It is not surprising that such an ambitious evaluation effort has engendered much discussion and debate on technical issues related to the reporting system. Concerns have covered such topics as test development and standardization procedures, test floor and ceiling effects, regression to the mean and the specification of a sound no-treatment expectation. A number of published and unpublished papers on these topics and others have been circulated.

The purpose of this paper is twofold. First, it will present an annotated bibliography of papers concerning technical issues surrounding the Title I Evaluation and Reporting System. Second, it will place the various issues within a unifying structure of threats to validity.

The framework for treating technical issues is based on Cook and Campbell's (1976) threats to validity as modified by Murray (1978). The framework includes "Five Types of Validity Relevant to Inferences Typically Made in Evaluative Research:"

I. Internal Validity:
The extent to which inferences of cause-effect relationships systematically and appropriately reject rival alternative causes.

II. Statistical Conclusion Validity:
The extent to which statistically-based conclusions about the existence or nonexistence of covariation are free from Type I and Type II errors.

III. External Validity:
The extent to which results may be expected to soundly generalize across persons, settings and times.

IV. Construct Validity of the Cause:
The extent to which operational definitions, either through manipulations or measurements, of independent variables are unambiguously interpretable at a given level of reduction.

V. Construct Validity of the Effect:
The extent to which operational definitions of dependent variables are unambiguously interpretable at a given level of reduction.

The Cook and Campbell framework was chosen because it is familiar to most educational evaluators. Numerous designs have been analyzed in terms of this framework and its precursor developed by Campbell & Stanley (1966). The literature on methodology includes many papers examining ways in which particular threats can interface with interpretations of results from different designs and ways in which specific threats can be controlled so that alternative explanations for results can be ruled out. Thus, the Cook and Campbell framework is used to provide a broader perspective for understanding the problems that could be encountered with each of the Title I evaluation models (described briefly in the next section). Tapping the existing literature could also help evaluators to validly implement designs and interpret results.

One potential problem with this classification scheme is the tendency to fragment issues. For example, one topic of special interest is the validity of the equipercentile assumption. There are several factors that affect the validity of the assumption, each of which could be classified as different threats. For example, the equipercentile assumption may be inaccurate because the maturational rates of Title I students are not the same as for the norm group (classified under maturation and selection), or because test levels are slightly misarticulated in the extremes of the distribution (classified as instrumentation), or because of statistical regression (classified under regression). Our rationale for proceeding with the Cook and Campbell framework is that first, only a few issues are fragmented in the process. More importantly, however, classification by threats instead of topics broadens the range of alternative strategies for resolving the issues. That is, while the topic of "validity of the equipercentile assumption" is empirically testable, the results of such examination tell one nothing about how to compensate or add controls to ensure the validity of the assumption.

We will begin by discussing the threats to validity that are related to the Evaluation System in general (that is, all three models). Following this, the threats to the validity of each model will be described. Comments will include the mechanisms used for control in a particular model (e.g., randomization as a control for selection in Model B), the adequacy of the control, and suggestions for other controls.

For details on the papers reviewed the reader will be directed to the annotated bibliography (see Appendix A). Each entry in the bibliography will be cross-referenced in two ways--first, to the topic it discusses (equipercentile assumption, norming problems, etc.), and second, to the threats to validity it represents.

Title I Evaluation and Reporting System Evaluation Models

The Title I Evaluation and Reporting System allows school districts to choose between three evaluation models. Each of the three can be used with normed or nonnormed tests, giving six options in all. The specially developed standard score, the NCE, has been introduced as a common metric for reporting project results. Each model and its nonnormed test option includes procedures for transforming treatment effects into the NCE metric. The models use different methods to generate a no-treatment expectation--an estimate of how project students would have performed without the project. The no-treatment expectation is compared to observed performance. The difference defines the treatment effect. It is claimed that if each model is implemented properly comparable results will be obtained. Thus, internal validity is an essential feature of each model.

Model A is called the norm-referenced model. It uses the norms of a test to generate the no-treatment expectation through the assumption that without supplemental help, students would maintain their percentile status over time. Thus the no-treatment expectation is the percentile equivalent of the average pretest score for the group. This is compared to posttest percentile status. Any change in status is said to be a measure of program impact.

Model B is the classic comparison group model. A comparison group is chosen that is as much like the Title I group as possible. The posttest performance of the comparison group (which may be adjusted using differences in pretest status) is then used as the no-treatment expectation. The difference between the Title I and comparison group's adjusted posttest scores is the measure of program impact.

Model C, the regression model, uses a group of students scoring higher than the treatment students on the pretest as the comparison students. It is assumed that without treatment the Title I students would retain their relative position with respect to the comparison students and hence would have scores that would tend to lie on the same post- on pretest regression line as that for the comparison group. Any difference between the post- on pretest regression lines for the Title I and comparison students would be a measure of program impact.

Threats to Validity Common to All Three Models

Because of the objectives and basic approach of the Title I Evaluation and Reporting System, some threats to validity are common to all three evaluation models. (The numbering system of validity threats corresponds to that in Appendix B.)

I. Internal Validity

D. Instrumentation

The Evaluation System relies on survey tests and normative scales for measuring achievement differences. As a result each of the models is vulnerable to instrumentation effects involving scaling. Theoretically, the Evaluation System is designed so that impact measures are comparable across tests, grades and pretest percentile status. In other words, a family of equal interval scales, NCEs, is assumed. Empirical findings, however, raise questions concerning this assumption. Some problems are related to norming practices (see Model A), others may be indicative of more theoretical problems with the academic growth construct as represented in the tests that claim to measure it.

1. Growth rates tend to vary across grades (Kaskowitz & Norwood, 1977; Norwood & Stearnes, 1976; Stenner et al, 1978; Tallmadge & Fagan, 1977), percentile status (Kaskowitz & Norwood, 1977; Norwood & Stearnes, 1976; Tallmadge & Fagan, 1977) and across tests (Stenner, et al, 1978). This means that a given gain may be harder or easier to detect depending on grade or test. (Percentile status may have more of an effect on Model A but also affects other models to the extent that treatment and comparison groups are nonsimilar as in Model C.) This detracts from making valid comparisons and aggregations across grades and tests.

2. Comparisons of student change across programs using different tests may not be meaningful because of differential sampling errors in the norms (Jaeger, undated; Tallmadge, 1977; Tallmadge & Horst, 1978), differences in content, differences in test/curriculum content match (Fishbein, 1978; Jenkins & Pany, 1978; Porter, et al, 1978; Tallmadge, 1977; Wiley, undated) or some combination of these factors.

3. The average of student NCEs may not equal the NCE of the average standard score. They will be the same only if the standard score to NCE conversion is linear (Norwood & Stearnes, 1976). Thus, computed results for groups may vary, which would affect comparisons.

4. Conversion to NCEs on nonnormed tests is linear while for normed tests it is based on the normal curve. The two conversion methods will be the same only when the underlying distribution of scores is normal. Since many nonnormed tests (CRTs) may have nonnormal distributions at pretest or posttest time, the two computation procedures would yield different results (Jaeger,

undated; Tallmadge, 1977). One study using normed tests (Deck, 1979) also showed differences between results using the two conversion methods.

5. If students are tested with a test that is too easy or too hard (Floor and Ceiling Effects), their true status and gain are under- or overestimated (Faddis and Estes, 1978). Observed results are thus not due to the project but rather due to the test level chosen.

Controls for general threats to internal validity (instrumentation):

1. Test at the functional level of students.

2. Always use area transformations to NCEs rather than linear transformations when it is possible to do so.

3. Convert individual scores to NCEs and do all calculations in terms of NCEs.

4. Match tests to curriculum.

II. Statistical Conclusion Validity

Since none of the evaluation models require tests of significance, the accuracy of statements concerning statistical significance is not a central issue. However, there are some problems that might affect Statistical Conclusion Validity if significance tests are used, and which in any case affect conclusions about the magnitude of a program's impact.

A. Statistical Power. Academic growth becomes more difficult to detect in the upper grades because learning decelerates at the same time that student scores are becoming more variable (and hence standard deviations are increasing). Therefore, normal growth is a decreasing fraction of the standard deviation of scores. Increasingly larger sample sizes are required in order to show significance in gains that are a decreasing fraction of the standard deviation. It takes a larger sample size to show significance in smaller relative increases. Or, to put it another way, it is harder to show significance in the higher grades with the same sample size (i.e., differences must be relatively large compared to normal growth) (Kaskowitz & Norwood, 1977; Norwood & Stearnes, 1976; Stenner et al, 1978; Tallmadge & Fagan, 1977). One implication of this is that the 1/3 standard deviation rule of thumb for significant change is easier to obtain in lower grades because 1/3 standard deviation may be 1/3 normal growth in lower grades and twice normal growth in the upper grades.

B. Fishing and the Error Rate Problem. This may be a problem in all models because gains are calculated for each grade within each school. Although only the aggregate is ever reported to the state departments of education, the calculation of so many gains may lead people to make multiple comparisons.

C. Reliability of Measures. The Title I Evaluation and Reporting System specifies that measures must be reliable. Rules of thumb (Ebel, 1965, p. 330; Helmstadter, 1964, p. 85; Halmstadter, 1970, pp. 294-295; Hills, 1976, p. 106; Nunnally, 1967, p. 173; Seibel, 1968, pp. 271-272):

1. When making decisions about individuals, one should have a reliability of at least .85 for examining status and .95 for measuring change.

2. When making decisions about groups, reliabilities can be lower—.65 for measuring status and .85 for measuring change.

D. Reliability of Treatment Implementation, E. Random Irrelevancies in the Experimental Setting and F. Random Heterogeneity of Respondents. Field evaluations of school programs are characteristically affected by lack of treatment standardization, heterogeneity of students and random irrelevancies. Although these effects are not attributable to the Evaluation System, their presence places a general limitation on the use of the Evaluation System.

Controls for threats to Statistical Conclusion Validity:

1. Make only comparisons that can be specified as important in advance of analysis.

2. If several comparisons are to be made, set to .01 or lower.

3. Describe interventions as clearly as possible and monitor implementation.

4. Observe rules of thumb for reliability.

III. External Validity. External validity is of little concern when using the Title I Evaluation and Reporting System when the purpose of the evaluation is confined to describing the effect of a program on the current group of students within the current year. There is little emphasis on using the Title I Evaluation and Reporting System to generalize to other students or school years. Generalizing from results is important in two situations: (1) In validating programs and (2) for making local decision about programs. In these cases threats to external validity should be examined:

A. Interaction of Treatments and Treatments. This may be a threat in Title I project evaluations because low-achieving students are often in a variety of other special programs such as bilingual or local remedial classes.

B. Interaction of Testing and Treatment. Results may be due to a particular testing situation--other situations would yield different results.

C. Interaction of Selection and Treatment. Can results be generalized to bilingual, inner city, rural or other specialized students? Is this important?

D. Interaction of Setting and Treatment. Can results be obtained in Title I programs operating in other school systems? This may be especially important in validating and disseminating.

E. Interaction of History and Treatment. Can results be obtained year after year? This issue is important in validating and decision making.

F. Generalizing across Effect-Constructs. Across what outcome measures can we expect effects? Cost-effectiveness decisions require answers to this question.

Controls for External Validity:

1. Replicate results across years, districts, outcome measures and populations.

2. Define programs clearly and investigate which program aspects are essential and which lead to various outcomes.

3. Make special note of unique aspects of project setting, staff and students. These features will reduce confidence in generalizations.

4. Examine missing data and its influence on results. Compare dropouts to nondropouts on pretest measures to see if results can be generalized to the whole population.

IV. Construct Validity of Cause. Problems in interpreting quasi-experimental results can arise from the way in which treatments are defined and operationalized. In general, evaluation does not concern itself with a match between the theoretical definition of a causal construct and its operational definition. If evaluation results are used to support interpretations of why the project is successful, however, construct validity of a cause does become an issue. A major problem is inadequate preoperational explication of the treatment construct (Linn, 1978).

V. Construct Validity of Effects

A. Inadequate Preoperational Definition of Effect Construct.

1. Poorly defined or nonexistent project objectives will detract from the valid interpretation of evaluation findings. Test selection cannot proceed rationally without objectives. Results could lack meaning and credibility if there is no basis for matching the content of a test with objectives or if there is a poor match between objectives and test content. Since the Evaluation System allows considerable flexibility in choosing a test, local implementation plays a crucial part in reducing this threat. It is noteworthy that descriptions of the Evaluation System provide little guidance in defining project objectives.

2. There is an additional factor that enters into the preoperational definition of achievement constructs. The construct of academic achievement seems to be undergoing considerable reshaping in recent years. In general, constructs are useful only insofar as they aid in understanding and predicting important phenomena. Traditionally, standardized achievement tests have focused on describing individual differences and prediction of future performance. Global achievement measures have served these purposes more efficiently than extremely specific measures. On the other hand, demands to evaluate large federal programs in terms of student achievement in basic skill areas has prompted a number of interpretation problems with standardized survey test data.

3. Lack of sensitivity to instruction is a major complaint against standardized norm-referenced achievement tests. Common sense dictates that a test used to measure instructional effects must be able to detect those effects if they exist. Edumetric validity is a term which has been coined to symbolize those properties of tests which give them sensitivity to instruction (Stenner, et al, 1977).

B. Mono-Operational Bias. This is a limitation of each of the models because in the Title I evaluation system emphasis is placed on academic achievement measured by a single test (Fishbein, 1978; Linn, 1978; Porter et al, 1979; Tallmadge, 1977). A more complete picture of program effectiveness would result if more than one method were used to measure achievement.

Controls for Construct Validity of Effects:

1. Write project objectives in measurable terms.

2. Match outcome measures to objectives.

3. Include more than one measure of each outcome and look for convergence across different measures of the same construct.

Model A: Threats to Validity

There are four points that must be taken into account in analyzing Model A within the Cook and Campbell framework. These points include:

1) The originality of Model A.

2) The absence of a formal statistical model underlying Model A.

3) The use of a normative, derived score as the common reporting metric.

4) The use of norms as a substitute control group.

Model A, because of its originality, as a quasi-experimental design, is easily the least studied of the three evaluation models making up the Evaluation System. The methodological development of Comparison group models (Model B) has been traced to the early part of this century (Campbell and Stanley, 1966). Refinements to this seemingly straightforward design are still being advanced (c.f., Cronbach et al., 1977). Model C, an extrapolation of the regression-discontinuity design introduced by Thistlethwaite and Campbell (1960) has a history of nearly twenty years. From this perspective Model A is a very recently proposed addition to the list of available quasi-experimental designs. Very little work on Model A has yet been published.

Model A is, in a fundamental sense, nonanalytic. That is, there is no explicit statistical model to support its application as a model for projecting unbiased no-treatment expectations. On the other hand, both Model B and Model C are based on formal statistical models with well known assumptions. Attempts to develop a statistical rationale for a norm-referenced model have resulted in approaches to generating unbiased no-treatment expectations that conflict with the originally proposed Model A (Echternacht, 1978b; Murray, 1978a). The absence of a formal statistical model contributes to ambiguity about Model A and prohibits definitive critique.

An analysis of Model A in terms of the Cook and Campbell framework must also take into account the use of norms as the basis for the evaluation metric and as a substitute for a comparison group. Many of the technical issues concerning Model A may be classified as threats resulting from

variations in instrumentation. These issues include the effects of sampling errors in norms, equating errors, interpolation and extrapolation errors and score transformation errors. Errors of these types can lead to biased results in specific project evaluations.

Additional threats to validity can be related to the use of norms as a pseudo comparison group. The pseudo comparison group is a subgroup of the norm group that scores at exactly the same level as the local Title I group on the pretest. Thus the pseudo comparison group is conceptualized as an equivalent comparison group. By simple assertion, the pseudo comparison group is assumed to maintain its percentile status from pretest time to posttest time. This is a very strong assumption, since the pseudo comparison group cannot, in actuality, be identified within the norm groups. Cross sectional norms may not even include the same individuals in the two norm samples used for pretest and posttest norms.

The pseudo comparison group is at the root of a number of technical issues concerning the use of norms in Model A. While Model B and Model C are not immune to issues regarding norms, Model A is unique in that norms are related to problems with scaling (i.e., instrumentation) and the control of alternative causes through appropriate comparison groups (e.g., maturation, testing, statistical regression, selection and interactions with selection).

These introductory remarks should help in understanding the emphasis that we give to Model A technical issues.

I. Internal Validity

A. History. According to Campbell & Stanley (1966, p. 5), history, as a threat to internal validity, is "the specific events occurring between the first and second measurement in addition to the experimental variable." In a strict comparison group design the only difference between what happens to the treatment and control groups is the intervention. Other differences (scheduling, major or minor disasters, etc.) may contribute to alternative explanations for outcomes. Where treatment and comparison groups are close--temporally as well as geographically--historical differences are minimized. In Model A since the comparison group is the norms which were developed at a different time and in different locations than any single group of Title I students, the amount of effect of historical happening is unknown. That is, since the norm group may not have been equally affected by extraneous events, the norms do not provide an adequate control for them. However, since these events are impossible to quantify, perhaps the evaluator should concentrate on more probable or controllable conditions.

B. Maturation. Students change over time simply as a result of getting older. Norms take this maturation into account at a gross level. In essence, norms account for the maturation of total populations or random samples of students from such populations. A random sample of students would be expected to score very near the 50th percentile on appropriate grade level norms. These same students would also be expected to score at the 50th percentile in some subsequent testing also given appropriate grade level norms. Norms do not necessarily accurately account for maturation of nonrandomly selected students. The equipercentile growth assumption, which is the Model A approach to modeling maturation, simply applies the general rule for total population maturation on normed tests to any selected subgroup of that population. That is, any selected group is expected to maintain its percentile status from pretest to posttest as long as it was not selected on the pretest. Maturation as a threat to validity is, therefore, closely related to selection and statistical regression as threats. As will be noted in the following discussions of selection statistical regression, a number of studies have seriously questioned the general application of Model A procedures for control over maturation and/or statistical regression.

C. Testing. There may be effects due to test taking a number of times. Students may become test-wise or get to know the test especially if the test is wide ranged, used year after year, or used for selection, pretesting and posttesting. Test scores may increase not because of the project but because the students know the test. This may be especially a problem in Model A because in the other models the same effects would tend to appear in the Comparison group as well as the treatment group and would, therefore, tend to cancel.

D. Instrumentation. Specific types of instrumentation problems include differential sampling errors in norms, errors resulting from scalar transformations, errors stemming from interpolation and extrapolation, errors of equating across levels and between forms and errors resulting from failure to adhere to standard test administration procedures including failure to test on norm dates. (See also the Instrumentation Section in General Threats.) All of these might result in an over- or underestimation of the no-treatment expectation (Armor, et al, 1976; Devito & Long, 1977; Noggle, 1977; Wood, 1978; Scherich, 1978; Slinde & Linn, 1977).

1. Sampling error in the national norms can create errors in Model A results (Jaeger, undated; Tallmadge, 1977; Tallmadge & Horst, 1978). Sampling error could be a biasing factor in the same testing series if the direction and magnitude of the error is not the same for all grades. This also causes a problem when comparing

results across tests for which sampling errors may be different. Several studies (Jaeger, undated; Wood, 1978) have demonstrated that such sampling errors are small but present and can bias individual project results.

2. If the equipercentile function (i.e., the function that plots the pretest standard score against the posttest standard score predicted by the equipercentile assumption) is not linear, then the unit of analysis (i.e., calculating on standard scores and then converting to NCEs vs. converting scores to NCEs first) may affect results (Wood, 1978). One study (Kaskowitz & Norwood, 1977) found the function to be fairly linear for the MAT. If the equipercentile assumption is not appropriate for disadvantaged populations, then the replacement function also needs to be linear.

3. When interpolation and/or extrapolation to test dates are necessary, different methods and assumptions may lead to different results (Baker & Williams, 1978; Bridgeman, 1978).

4. Smoothing and fitting of empirical data occurs when scores are equated across forms and levels of a test (as when developing expanded scale scores). These errors tend to be small but add up in the extremes of the distribution and when going out of level and across forms (Crowder, 1978; Fuentes & Jeffries, undated; Haenn & Proctor, 1978; Hiscox & Owens, 1978; Long, et al, 1977; Ozenne, 1978; Powers, 1978; Slaughter & Gulles, 1978; Slinde & Linn, 1977; Wood, 1977).

   When Model A2 is used, there is evidence that equating the nonnormed test to a normed test can yield different estimates of impact than if the nonnormed test had norms of its own (Linn, 1978).

5. Fall-fall (or spring-spring) evaluations routinely give different measures of impact than fall-spring evaluation (David & Pelavin, 1978; Devito & Long, 1977; Faddis & Estes, 1978; Hardy, 1979; Holthouse, et al, 1976; South Carolina Evaluation Report, 1975). This could be interpreted as a result of differences between the normal maturational process of Title I and regular students (regular students continue to gain when not in school while Title I students do not). Or it may be interpreted as a scaling artifact resulting from assumptions about summer growth in the norms (i.e., some students appear to lose ground because an expectation for summer growth is built into the norms).

E. Statistical Regression. Evaluation designs applied to extreme groups such as Title I students are often prone to bias from statistical regression, sometimes called "regression to the mean." When retested, groups selected for their extremity tend to score closer to the mean of the population from which they were selected. Estimates of the regression effect in Model A show that when selection is on the pretest, regression is considerable (Mandeville, 1978; Yap, 1978).

The "control" against regression effects in Model A is to use one measure to select students and a second measure as a pretest from which a no-treatment expectation is derived. The assumption is that there is no further statistical regression from the pretest to the posttest. However, there have been several demonstrations that this procedure eliminates regression effects only under very unusual circumstances (Burton, 1978; Glass, 1978; Hiscox & Owens, 1978; Murray, 1978a). Campbell and Stanley (1966) cautioned researchers about this procedure. Their comment is instructive.

> Regression effects of a more indirect sort can be due to selection of extreme scorers on measures other than the pretest. Consider a case in which students who "fail" a classroom examination are selected for experimental coaching. As a pretest, Form A of a standard achievement test is given, and as a posttest, Form B. It is probable that the classroom test correlates more highly with the immediate Form A administration than with the Form B administration some three months later (if the test had been given to the whole class on each occasion). The higher the correlation, the less regression toward the mean. Thus the classroom failures will have regressed upward less on the pretest than on the posttest, providing a pseudogain which might have been mistaken for a successful remedial-education effort. (p. 12)

It has been demonstrated, however, that although Model A does not necessarily eliminate bias due to statistical regression it will most likely reduce such a threat (Glass, 1978; Murray, 1978a). To put this into proper perspective it must be remembered that even conventional analyses (MANCOVA) which are often used to provide an ex post facto statistical control for regression can underestimate as well as overestimate treatment effects (Cronbach et al, 1977). The problem creating these conclusions is differential regression. When differential regression is operating, maturational differences between subgroups as well as random errors of measurement are contributing factors to bias. Thus, it is probably fair to say that an exacting generalizable control applied in an ex post facto analysis is not possible.

F. Selection. Errors in estimation of impact can occur if the students in the treatment and comparison groups were not similar to begin with. This results in the comparison group not providing a good no-treatment expectation against which to compare the progress of the treatment students. When a real comparison group is available and the procedure for identifying the treatment and comparison groups is clear, it is possible to appraise the extent of selection effects. Since Model A uses a pseudo comparison group it is difficult to form any strong hypotheses concerning the extent to which Model A controls for selection. Because the pseudo comparison group concept is used to further a maturational hypothesis (i.e., groups maintain their percentile status from time to time), selection effects can be combined with maturation in a selection x maturation interaction. Other interactions with selection are also possible. These will be considered in the section on interactions.

G. Mortality. Mortality could be a problem in Model A insofar as there is differential mortality: different types of students dropping out of the Title I group and norm groups. Since this is an interaction with selection it will be considered in the next section.

H. Interactions with Selection.

1. Mortality and Selection. Low students may tend to drop out of school over time (especially in high school). This might result in a norm group that is getting progressively "better." If there is a differential mortality between the regular and special programs (that is, if a special program is keeping students in school) the treatment group might tend to fall off in percentile over time. One study that examined this issue failed to find support for the hypothesis (Arter & Estes, 1978).

2. Instrumentation and Selection.

   a. The learning and maturation curves for local treatment students may not be the same as those for the norm group. This implies a selection x maturation effect. Instead of maintaining percentile status, Title I students may indeed fall progressively further behind with time (Linn, 1978; Noggle, 1977). Several studies have been undertaken to determine whether this occurs (Coleman, et al, 1976; Doherty, undated; Faddis & Estes, 1978; Hiscox & Owens, 1978; Kaskowitz & Norwood, 1977; Mayeske & Beaton, 1975; Norwood & Stearnes, 1976; Van Hove, et al, 1970). Results have been mixed probably due to methodological and population differences.

b. Publishers generally develop cross-sectional norms (norming in all grades at the same time) rather than longitudinal norms (testing the same students as they progress through grades). The result is that even though each grade sample represents the national population at that grade, the students in each grade are different. Thus, the pre and post comparison groups for a program may be more or less similar to the constant Title I group (Hiscox & Owens, 1978; Kaskowitz & Norwood, 1977). Estimates of percentile status may change merely because the norm group is changing. One study on the MAT (Kaskowitz & Norwood, 1977) showed a fairly good fit between cross-sectional and longitudinal norms.

3. Testing and Selection. Testing at times other than norm dates and not following instructions exactly can also result in errors. Since the norm group acts as a pseudo comparison group in Model A, testing off norm dates makes incorrect comparisons of status. Not following test instructions results in noncomparable testing conditions, rendering comparisons of status ambiguous.

I. Diffusion or Imitation of the Treatment. Excluding Title I students from norming procedures is not a standard practice. Thus treatment effects are included in the norms, resulting in impact estimates that can be artificially low.

K. Compensatory Equalization of the Treatment, L. Compensatory Rivalry and M. Resentful Demoralization of Respondents Receiving Less Desirable Treatments. These factors seem very remote since the comparison group is not local.

N. Local History. See History.

Controls for Threats to Internal Validity:

1. Use different forms of a test if usages are fairly close together or if the same students are likely to be in a program several years, and well hence be seeing the same test many times. This eliminates testing threats.

2. Use the same level and form of a test pre and post as much as possible. This reduces error due to scaling. However, error due to repeated test administrations or incorrect test levels might outweigh errors of scaling. If so, then you would reduce error by changing test forms or levels.

3. Check test manuals for estimates of sampling error to make sure that it is consistent in magnitude at each grade level.

4. Check test manuals to make sure that errors of equating are small.

5. Separate fall-fall and fall-spring test data.

6. Test at empirical norm dates and follow test instructions exactly.

7. To reduce regression effects, separate the selection and pretest measures. More sophisticated controls can be implemented depending on circumstances:

   a. Check whether $r_{12} = r_{13}$, where 1 = selection test, 2 = pretest and 3 = posttest. If the correlations are equal, then all regression has been eliminated by the separation of selection and pretesting. (This is not likely.)

   b. If a strict cut-off has been maintained on the selection measure, it can also be used as the pretest. The no-treatment expectation is then

   $$\bar{x}_2 = \bar{x}_1 r_{12}$$

   (Echternacht, 1978b; Murray, 1978a). This design is very similar to Model C. The only difference is that population estimates are supplied by the norm group instead of the local group.

   c. Yap (1978) uses another estimation procedure for estimating the no-treatment expectation.

8. Controls for selection by maturation interactions. Some ways in which a local district may check whether growth rates are the same for their students as for the students in the norms are:

   a. Go back to a time when there was no Title I program in the district. Choose students who would have been classified as Title I had there been a program and trace their percentile status over a period of time matching the amount of time between a pretest and posttest date. If it remains constant, then see whether entering students today are similar to those in the past (pretest mean, racial composition, etc.). If so, then the equipercentile assumption may hold up reasonably well.

   b. Trace the status of the district as a whole over time to see if it maintains its percentile status. If so, it might be assumed that Title I students (or other subpopulations) would too. This method does not require as much historical data as in (a) but is not as accurate.

c. Identify a local group of students similar to Title I students but not in Title I. See whether they maintain status over time.

9. Because of the absence of norms for students like Title I students untreated by Title I programs, some evaluators have suggested developing Title I norms (Committee on Model A, 1978). Progress of students would not be compared to no-treatment but to the cumulative effects of treatment in the norming year.

II. Statistical Conclusion Validity. As pointed out in the section on general threats to all three models, no significance testing occurs in the Title I Evaluation and Reporting System. If statistical tests were used, however, Model A may be particularly susceptible to lack of Statistical Power since programs with small sample sizes often choose Model A as a default.

III. External Validity. See General Threats

IV. Construct Validity of Cause. See General Threats.

V. Construct Validity of Effects. See General Threats.

Model B: Threats to Validity

I. Internal Validity. When implemented as intended, Model B is relatively free from threats to internal validity. Its properties have been discussed at length by Campbell and Stanley (1966) and Cook and Campbell (1976). The reader is referred to these sources for more thorough analyses. Note that Campbell and Stanley (1966) discuss two designs (i.e., the Pretest-Posttest Control Group Design and the Nonequivalent Control Group Design) which are included as special variations of Model B in the USOE Title I Evaluation and Reporting System.

The ratings in Table 1 assume that a reasonable comparison group has been found and that the appropriate method of analysis (i.e., no adjustment, analysis of covariance adjustment or principle axis adjustment) has been selected. These ratings should not be assumed to hold for any specific application of Model B. Breakdowns in implementation can reintroduce specific threats. Breakdowns could result from such things as unclear guidelines, poor understanding of the guidelines, lack of information about conditions affecting the tenability of assumptions made in Model B and compliance oriented implementation.

The following discussion will focus on some common threats to validity when Model B cannot be properly implemented. Most of these problems occur because of the difficulty of setting up good control groups in a natural setting.

A. History, C. Testing. These should present little difficulty even if selection is not optimum because both treatment and control groups are local. These will only be a problem if they interact with selection because of some atypical arrangement (see section H).

B. Maturation, E. Statistical Regression, G. Mortality. These should not be a problem if selection is reasonable. If selection is not reasonable, then these items may affect results to the extent that they are different between groups (see section H).

D. Instrumentation. When a nonnormed test is used in Models B or C, extra computations are necessary (using the RMC framework) in order to convert results to NCEs. The first step is to estimate the standard deviation of the national population on the nonnormed test using the formula:

$$\frac{SD_{N,NN}}{SD_{L,NN}} = \frac{SD_{N,NRT}}{SD_{L,NRT}}$$

NN = nonnormed test
NRT = normed test
N = national population
L=local population

or $$SD_{N,NN} = \frac{SD_{N,NRT} \times SD_{L,NN}}{SD_{L,NRT}}$$

Then $SD_{N,NN}$ is used to make a linear conversion to NCEs:

$$\text{NCE gain} = \frac{21.06}{SD_{N,NN}} (T_2 - C_2)$$

T_2=X for Title I group on posttest
C_2=X for Comparison group on posttest

The second computation was discussed in the computation section of General Threats. The first conversion is of interest here. There is some evidence that $SD_{N,NN}$, as computed by the ratio described above, can be very misleading (Fishbein, 1978; Long, et al, 1978).

F. Selection. Ideally, selection should proceed through random assignment. Almost by definition this would be illegal since the lowest students must be served in Title I.

1. Close approximations to the ideal design can occur when: (a) Students are in the program for one semester and only half the students are served at any one time. Students may be randomly assigned to Title I or Comparison groups each semester, and the evaluation could occur only for the first semester. (b) If only enough money exists for some eligible schools to be Title I schools, then the others could be comparison schools (pay particular attention to differential regression and comparability of regular school programs). (c) If some students are

eligible for Title I classes but do not enroll (as at the high school level) they may be a comparison group, if there is no reason to believe that the reasons they did not enroll make them noncomparable to treatment students. For example, voluntary enrollment might yield different types of students while scheduling problems might not. If groups are random or random-in-effect (Tallmadge & Wood, 1976), Model B recommends covariance or no adjustment depending on the difference between the groups on the pretest. There has been some controversy over the legitimacy of the random-in-effect design.

2. Nonequivalent control group. Model B also provides a design when groups are assumed not to be the same (i.e., assumed to hail from different populations). This method adjusts posttest scores using a standardized gain score technique called the Principal Axis (PA) Adjustment. This assumes (as in Model A) that without treatment, students will maintain their relative positions through time. If this fan spread hypothesis is false, then the PA Adjustment may be biased (Hiscox & Owens, 1978; Holthouse, et al, 1978; Kaskowitz & Norwood, 1977; Linn, 1978; Norwood & Stearnes, 1976). Differential maturation, mortality and regression may especially affect these results.

H. Interactions with Selection. When selection is not random many interactions (i.e., differential effects for one group over another) may occur. Some examples are:

1. Practice in test taking may affect some groups more than others.

2. Instrumentation may be a problem if groups are not measured at the same time, or if a test measures one curriculum better than another.

3. Testing might be a problem if different forms or levels are used for the treatment and comparison groups (Wood, 1977).

4. Maturation for the two groups may not be the same, and therefore the comparison group scores may not yield a good no-treatment expectation for the Title I students.

5. Another example is differential regression (Tallmadge & Wood, 1976). If students are selected to be Title I or Comparison in different schools or classes based on a common cut-off percentile, then one group may be farther from the mean of its population than the other and hence will tend to regress more.

J. Diffusion or Imitation of the Treatment, K. Compensatory Equalization of the Treatment, L. Compensatory Rivalry, M. Resentful Demoralization of Respondents Using Less Desirable Treatments, N. Local History. All of these are more possible in Model B than in Model A since comparison groups are local. Whether or not they occur is more a function of the project than of the models.

Controls for Threats to Internal Validity

1. Controls to check on similarity of groups:

   a. If groups were not randomly assigned, check to see if groups are random-in-effect (Tallmadge & Wood, 1976), i.e., are similar in terms of race, sex, grade, SES, pretest measures, attitudes, etc.

   b. Look at historical data to see if dropout rates, learning curves, etc., are similar for groups that would have been Title I & Comparison had there been a Title I program.

2. Controls to avoid differential regression: Select the lowest X% of students in each school rather than selecting all students below a certain percentile (Tallmadge & Wood, 1976), or look at historical data (Echternacht, 1978a).

3. Controls for diffusion of treatment, compensatory rivalry, etc.: Program evaluators could monitor their programs to observe the extent of these occurrences using:
   a. Classroom visitation

   b. Interviews of students and teachers

   c. Questionnaires of students and teachers

II. Statistical Conclusion Validity. See General Threats.

III. External Validity. See General Threats.

IV. Construct Validity of Cause. See General Threats. One extra problem might be the requirement to have the treatment and comparison groups have the same instruction except for the treatment. This is probably not reasonable or possible.

V. Construct Validity of Effects. See General Threats.

Model C: Threats to Validity

Model C, the special regression model, requires that students be selected into the Title I program on the basis of a designated covariate only. The covariate can be a pretest or a composite score including or excluding a pretest score. It is assumed that comparison group students and selected Title I students are members of a single population. Under this assumption a linear regression equation predicting posttest scores from the covariate is derived from comparison group students who have not received supplemental services. This regression equation is applied to the Title I group. The result is a predicted posttest score (i.e., a no-treatment expectation) for the Title I group.

I. Internal Validity

A. History. See Model B.

B Maturation. The posttest on pretest linear regression takes into account changes in the general level of performance from the pretest to the posttest. It is in this way that general maturation is controlled in Model C. Selection x maturation interactions, as discussed under interactions with selection, can pose a threat, however.

C. Testing. Since both groups experience pretesting and posttesting, this threat to validity is controlled.

D. Instrumentation.

1. Floor and ceiling effects can affect the regression lines used to estimate the no-treatment expectation (Estes & Anderson, 1978).

2. In order to convert evaluation results to NCEs, the standard deviation of the national population must be estimated using the ratio of standard deviations of the treatment group on the normed and nonnormed tests. See the Instrumentation section under Model B.

E. Statistical Regression. Model C requires that Title I students be selected only on the pretest or a composite score. This implies a strict cutoff score. Given that the guideline is followed, the process for selection is determined and the effects of regression are taken into account in the analyses. Currently, work is continuing on the effects of relaxing the requirement that selection be based only on an available covariate.

F. Selection. Even though students of different ability are in the treatment and comparison groups, effects of these prior differences are controlled by selection on a covariate that assumes comparison groups and the Title I group are drawn from a single population and that the relationship between the covariate and the posttest is linear.

G. Mortality. Model C requires that both pretest (or a composite score) and posttest scores are available to complete the analysis. Students who drop out of either the treatment group or the comparison group could effect the results of a Model C analysis.

H. Interactions with Selection. The possible interactions which could pose threats to the internal validity of Model C include selection x maturation, selection x testing, selection x statistical regression, and selection x mortality. The presence of nonlinear regression or heterogeneous regression could be interpreted as evidence of such interactions. Additional evidence of a more substantive nature would be needed to specify the causes of a nonlinear or heterogeneous regression.

J. Diffusion or Imitation of the Treatment, K. Compensatory Equalization of the Treatment, L. Compensatory Rivalry and M. Resentful Demoralization of Respondents Using Less Desirable Treatments. All of these are more possible in Model C than in Model A since comparison groups are local. Whether or not they occur is more a function of implementation than of the models.

N. Local History. This can always be a problem in Model C. The best thing to do is to monitor what goes on and note any unusual occurrences.

Controls for threats to Internal Validity:

1. To see whether floor or ceiling effects are present in data, plot them and see if there is a buildup of scores around chance level or near the top score. The common remedy for reducing floor or ceiling effects is to test students at their functional level. However, this might increase the chances of selection by testing interactions or other instrumentation threats to validity.

   If a floor or ceiling is present, one might drop out students with chance level or top scores and redo analyses with those students who are left. If this is done, however, remember that results may not be generalizable to the entire population.

2. Evaluators should be sensitive to the factors of diffusion of treatment, compensatory equalization, compensatory rivalry and local history, and use such techniques as classroom visitation, interviews of students and teachers and questionnaires of students and teachers to assess their presence.

3. Heterogeneous regressions might indicate floor and ceiling effects or may indicate other interactions with selection.

4. To examine the effects of mortality, one could compare the pretest scores and background characteristics of students who remained in the Title I and comparison groups and those who did not. If there are no differences it is more likely that results could be generalized to the whole population. Differences may result in selection x mortality interactions.

II. Statistical Conclusions Validity.

A. Power. Model C has a larger standard error than Model B. Therefore, more students are required to demonstrate the same effects.

III. External Validity. See General Threats.

IV. Construct Validity of Cause.

It may be unlikely that Title I and comparison students have the same instruction except for treatment. Therefore, evaluation results may not be totally attributable to the Title I program.

V. Construct Validity of Effects. See Model B.

Conclusions

In the process of reviewing papers and categorizing Title I Evaluation System Issues by threats to validity, several considerations arose that should be examined. These considerations relate to confounding or confusing conclusions in papers, relative emphasis of study types and ambiguity about the purposes of a study. Hopefully, a discussion of these problems will lead to clearer conceptualization of research and suggest clearer research directions.

The first consideration has been partially discussed already--the fragmentation of issues when considering validity threats, and the fragmentation of validity threats when considering technical issues. For some issues and threats there is a direct relationship. One example is regression. The most common issues that lend to fragmentation are the validity of the equalpercentile assumption, summer loss, spring-spring versus fall-spring evaluation cycles and problems in aggregating across tests, grades and initial percentiles.

There are so many things that can influence studies on these topics that it is often difficult to sort out and control for all but the variable of interest. Thus many studies tend to show inconsistent results and conclusions that do not account for other interpretations.

A second consideration is that many studies relate to validity threats within a single evaluation while other studies deal with the validity of aggregating results and still others have effects on both. Examples of the threats to validity of aggregation are differing growth rates across grades and initial percentile status, differential sampling errors in

various tests, analyzing results, using students or classes as the unit of analysis, linear conversion to NCEs on nonnormed tests versus area transformations on normed tests, differential floor and ceiling effects, using different models and differential test to curriculum matches. Besides questions of how to classify these threats, there are also other problems. One is the question of the definition of "comparable." Some would say that "comparable" means getting the same result using two different tests or models (Jaeger, undated). Others say that when tests and models are chosen that are most appropriate for a program, the results are comparable even when other tests or models yield different results (Tallmadge, 1978).

A second problem is whether all the errors tend to average out when results are aggregated, and whether there is so much error variance in the results that time effects are masked.

A third problem relates to conclusions regarding the construct of achievement based on tests. To illustrate this point, consider the attempt to verify the equalpercentile assumption. It is empirically possible to measure status of individuals or groups over time using a test and to see if percentile standing remains constant. It does not logically follow, however, that results of such empirical studies can be used to advance conclusions about the construct of achievement. For example, if percentile status falls over time for minority students, it is tempting to say that minority students learn at different rates. However, the results could also be attributed to the manner in which the test was constructed (cross-sectional or nonempirical norms assumptions concerning summer growth). In other words, a construct cannot be separated from the tests that claim to measure it without multiple lines of evidence.

A final concern deals with a paint-by-number approach to evaluation. The question is whether evaluation is at such a state that valid results and conclusions can be made by persons not practiced in evaluation. Some believe that interpretations and conclusions must be made in light of design problems, data collections and analysis problems, measurement problems and program knowledge.

The Title I Evaluation and Reporting System is the first nationwide attempt to standardize evaluation practice and obtain comparable evaluation results across programs. As such, it is an admirable attempt to combine theoretical rigor with practical necessity. The Title I Evaluation and Reporting System has had a broad impact on the field of evaluation. Spinoffs have been noted in at least four areas: evaluation practices and sophistication at the local level, legislation and Congressional actions, effects on the field of evaluation and test development (Arter & Estes, 1978).

We must conclude, however, by saying that it is clear that none of the Title I Evaluation and Reporting System evaluation models is without threats to validity. Models B and C are rigorous but are difficult to

implement properly. Model A has the most theoretical problems, but may be useful if data checks are used. Indeed, the most important aspect when using any of the models is not the inherent goodness of the model itself, but the care that is taken when designing the evaluation, analyzing the data and interpreting the results.

Table 1
Summary of the Threats to the Validity of the Various
Title I Evaluation and Reporting System Evaluation Models

Threat/Model	A	B	C
I. Internal Validity			
A. History	-	+	+
B. Maturation	0	+	0
C. Testing	0	+	+
D. Instrumentation	-	+	+
E. Statistical Regression	0	+	+
F. Selection	-	+	-
G. Mortality	-	+	-
H. Interactions with Selection	0	+	0
I. Ambiguity about Direction	+	+	+
J. Diffusion of Treatment	-	+	+
K. Compensatory Equalization	+	0	0
L. Compensatory Rivalry	+	0	0
M. Resentful Demoralization	+	0	0
N. Local History	-	+	+
II. Statistical Conclusion Validity			
A. Statistical Power	-	-	-
B. Error Rate	-	-	-
C. Realiability of Measure	+	+	+
D. Reliability of Treatment Implementation	-	-	-
E. Random Irrelevancies	-	-	-
F. Random Heterogeneity	-	-	-
III. External Validity			
A. Interaction of T & T	-	-	-
B. Interaction of T & T Testing	-	-	-
C. Interaction of T & T Selection	-	-	-
D. Interaction of T & T Setting	-	-	-
E. Interaction of T & T History	-	-	-
F. Generalizing Across Effect Constructs	-	-	-

+ = Controlled

0 = Controlled if precautions are taken

- = Uncontrolled

Threat/Model	A	B	C
IV. Construct Validity of Cause			
A. Inadequate Explication of Treatment	0	0	0
B. Mono-Operational Bias	-	-	-
C. Mono-Method Bias	-	-	-
D. Hypothesis Guessing	0	0	0
E. Evaluation Apprehension	0	0	0
F. Experimenters Expectancies	0	0	0
G. Confounding Treatment	0	0	0
H. Interaction of T & Procedure	-	-	-
V. Construct Validity of Effects			
A. Inadequate Explication of Effects	-	-	-
B. Mono-Operational Bias	-	-	-
C. Mono-Method Bias	-	-	-
D. Hypothesis Guessing	0	0	0
E. Evaluation Apprehension	0	0	0
F. Experimenter Expectancies	0	0	0
G. Confounding Levels	0	0	0
H. Generalizing Across Time	-	-	-

REFERENCES

Arter, J. A., and Estes, G. D. Spinoffs from the USOE Title I evaluation and reporting system. TACKY Issue Paper #10, NWREL/TAC, 1978.

Campbell, P. T. and Stanley, J. C. Experimental and quazi-experimental designs for research. Chicago: Rand McNally, 1966.

Cook, T. D., and Camp ell, D. T. The design and conduct of quasi-experiments and true experiments in field settings. M. D. Dunnette (Ed.) Handbook of Industrial and Organizational Psychology. Chicago: Rand McNally, 1976.

Cronbach, L.J.; Rogosa, D. R.; Ploden, R. E.; and Price, G. G. Analysis of of covariance in nonrandomized experiments: Parameters affecting bias. Stanford Eval uation Consortium, Occasional Paper, 1976.

Ebel, R. L. Measuring educational achievement. Prentice-Hall, Inc: Englewood Cliffs, New Jersey, 1965.

Helmstadter, G. C. Principles of psychological measurement. Appleton-Century-Crofts: New York, 1964.

Helmstadter, G. C. Research concepts in human behavior. Appleton-Century- Crofts: New York, 1970.

Hills, J. R. Measurement and evaluation in the classroom. Charles E. Merrill Publishing Co. : Columbus, Ohio, 1976.

Nunnally, J. C. Psychometric theory. McGraw-Hill Book Co.: New York, 1967.

Seibel, P. W. Measurement of aptitude and achievement. In p.k. whitla (Ed.) Handbook of measurement and assessment in behavioral sciences. Wesley Publishing Co.: Reading, Massachusetts, 1968.

Tallmadge, G. K., and Wood, C. T. User's guide: ESEA Title I evaluation and reporting system. Mountain View, California: RMC Research Corp., 1976.

Thistlethwaite, D. L., and Campbell, D. T. Regression-discontinuity analysis: An alternative to the ex post facto experiment. Journal of Educational Psychology, 1960.

APPENDIX B

Five Types of Validity Relevant to Inferences Typically Made in Evaluation Research

I. Internal Validity -

The internal validity of an experiment or quasi-experiment is the degree to which alternative plausible explanations for changes in a dependent variable have been eliminated because of design features.

Common threats to internal validity:

A. History: events that are not a part of the "treatment" and which take place between a pretest and a posttest may impact upon the dependent variable.

B. Maturation: spontaneous or natural development of a respondent during the treatment.

C. Testing: effects due to test-taking a number of times.

D. Instrumentation: effects due to changes in the measuring instrument.

E. Statistical Regression: effects due to the selection of extreme groups as experimentals or controls.

F. Selection: effects due to differences in the kinds of persons in experimental groups (prior).

G. Mortality: effects due to subjects dropping out of the experiment before its completion.

H. Interactions with Selection: effects due to the interaction of history, maturation, mortality, etc. with selection.

I. Ambiguity about the Direction of Influence: effects of A and B are clouded by research that does not clearly allow for testing of the direction of causality.

J. Diffusion or Imitation of the Treatment: effects of the control group's adoption of procedures that are also part of the treatment--the impact of this is to decrease differences between groups.

K. Compensatory Equalization of the Treatment: effects of adding resources to control groups as an administrative or political device for providing equity--the impact of this would be to decrease differences.

L. Compensatory Rivalry: competition effects resulting from public declaration of membership in either experimental or control groups.

M. Resentful Demoralization of Respondents Receiving Less Desirable Treatments: differences due to response of resentment on the part of respondents.

N. Local History: effects of group-specific events in treatment or testing.

II. Statistical Conclusion Validity

The statistical conclusion validity of a finding is the extent to which conclusions about covariation are true. It is a special case of internal validity.

Common threats to statistical conclusion validity:

A. Statistical Power: the likelihood of making an incorrect no-difference conclusion (Type II error) increases with small sample sizes, low x, incorrect one-tailed hypotheses and the use of distribution-free statistics for hypothesis testing.

B. Fishing and the Error Rate Problem: Type I errors will result from making multiple comparisons of mean differences.

C. Reliability of Measures: Type II errors will be increased by unreliability of measures.

D. Reliability of Treatment Implementation: Type II errors will be increased by a lack of treatment standardization.

E. Random Irrelevancies in the Experimental Setting: will increase Type II errors (see interaction of procedures and treatment).

F. Random Heterogeneity of Respondents: the greater the heterogeneity of experimental subjects on variables that covary with the dependent variable, the greater the chances of Type II errors.

III. External Validity

The degree to which results are generalizable across times, settings and persons indicates the external validity of an experiment. External validity is related to but distinct from the concept of construct validity, where the issue is the level of generalization across different effect constructs (e.g., learning an alphabet and reading). Snow (1974) has referred to "referent generality" to apply to the multiple-outcome problem.

Common threats to external validity:

A. Interaction of Treatments and Treatments: where subjects have received more than one treatment, we do not know that results will generalize to settings where only one treatment is applied.

B. Interaction of Testing and Treatment: to what kinds of testing situations can a cause-effect relationship be generalized?

C. Interaction of Selection and Treatment: across what categories of people can cause-effect relationships be generalized?

D. Interaction of Setting and Treatment: can causal relationships be obtained in a variety of school districts?

E. Interaction of History and Treatment: to which periods--past and future--can a particular causal relationship be generalized?

F. Generalizing across Effect-Constructs: can results of experimentation be generalized across different outcome measures?

IV. Construct Validity of Cause

The construct validity of a cause is an evaluation of the inference linking manipulations and the abstract construct which it represents.

Common threats to construct validity of cause: construct under-representation, construct vagueness, construct surplus irrelevancies.

A. Inadequate Preoperational Explication of the Treatment Construct: poor or otherwise insufficient behavioral or procedural definition of the treatment is supplied prior to the initiation of the experiment ("vagueness").

B. Mono-Operational Bias: distortions from reliance upon single means of presenting treatments ("under-representation").

C. Mono-Method Bias: distortions resulting from reliance upon single means of presenting treatments ("under-representation").

D. Hypothesis Guessing within Experimental Conditions: confounding of subjects' perceptions of hypothesized or expected results of exposure to experimental conditions with the experimental conditions themselves ("irrelevancy").

E. Evaluation Apprehension: fear or anxiety over the thought of being evaluated as it may confound or interact with the treatment ("irrelevancy").

F. Experimenter Expectancies: views of probably experimental outcomes held by the experimenter.

G. Confounding Levels of Treatment with Treatment: confounding discrete levels of continuous distributed treatment with the treatment, i.e., strength of treatment.

H. Interaction of Procedure and Treatment: effects of the context in which the treatment is administered as it impacts upon the perceptions of participants.

V. Construct Validity of Effects

Construct validity of effects involves an evaluation of measures that represent the abstract effect constructs.

Common threats to construct validity of effects:

A. Inadequate Preoperational Explication of Effect Constructs: poor definition of the effect constructs is supplied prior to measurement.

B. Mono-Operational Bias: distortions resulting from single measures of effects.

C. Mono-Method Bias: distortions resulting from reliance upon single means of recording responses.

D. Hypothesis--Guessing within Experimental Conditions: the occurrence of subjects' determining the predicted or expected results of treatments and responding to test items in accordance with those expected results.

E. Evaluation Apprehension: a state of apprehension over being evaluated by "experts" as it impacts upon responses to "test" items.

F. Experimenter Expectancies: instances of experimenter's holding expectations about the probable effects of a treatment.

G. Confounding Levels of Effect Constructs and Effect Constructs: confounding discrete levels of effect (which is continuously distributed) with "the effect."

H. Generalizing across Time: stability of effects, timing of effects, etc. (might best be under external validity)

APPENDIX A

Annotated Bibliography

Format of Bibliography

Reference

Text of Annotation

Editorial Comments

Technical Issues	Threats to Validity
Classification categories are presented on the next page	Reflect placement in the text and are numbered according to Appendix B

Technical Issues Categories

1. Norming Issues
   a. Sampling Error
   b. Vertical Scaling (Equating)
   c. Horizontal Scaling (Equating)
   d. Empirical Norms
   e. Norming Dates
   f. Estimates of Summer Growth
   g. Cross-Sectional v. Longitudinal Norms
   h. Contamination of Norms

2. Aggregation Problems
   a. Across Tests
   b. Across Grades
   c. Across Beginning Percentile Status
   d. Across Testing Cycle (Fall-Spring v. Spring-Spring)

3. Nonnormed Test Problems
   a. Linear v. Area Transformation to NCEs
   b. Estimates of SD of National Population
   c. Equating Nonnormed Test to NRT
   d. Floor and Ceiling Effects

4. Model Selection/Bias
   a. Comparison of Models
   b. Simulations on Bias
   c. Alternative Models
   d. When to Choose Various Models

5. Computation Problems
   a. Unit of Analysis
   b. Interpolation and Extrapolation

6. Regression
   a. Estimates of
   b. Corrections for

7. Sustained Effects
   a. Fall-Spring v. Spring-Spring
   b. Summer Loss

8. Equipercentile Assumption

9. Treatment Definition

10. Problems with NCEs

11. Test/Curriculum Match

12. Procedures for Out-of-Level Testing

Armor, D., Conry-Oseguera, P., Cox, M., King, N., McDonnell, L., Pascal, A., Pauly, E., & Zellman, G. Analysis of the school preferred reading program in selected Los Angeles minority schools. Report prepared for the Los Angeles Unified School District, R-2007-LAUSD, Rand Corporation, Santa Monica, California, 1976. (Also ERIC 130 243.)

The authors report on an evaluation of the Preferred Reading Program which attempted to relate background and school variables of students to gains in achievement. One part of this study traced the achievement patterns of 759 predominantly minority program students through four years. Results showed a percentile increase between grades 3 and 4 and steady percentile decreases thereafter (although raw scores continued to increase). The anomaly in the first year is attributed to a change in tests. The decrease is from the 33rd percentile in grade 4 to the 29th percentile in grade 6.

Comments:

1. Confounded with any real loss in percentile status is treatment effects. The equipercentile assumption states "in the absence of special programs." Most of the students in this study were in special programs. Thus losses could be interpreted as a measure of program impact.

Equipercentile Assumption

Model A
I. Internal Validity
D. Instrumentation

Arter, J. A., and Estes, G. D. A model for developing local norms with a standardized achievement measure for use with local program evaluation. Paper presented at AERA, Toronto, 1978.

The authors feel that one reason a norm-referenced evaluation may show inaccurate gains over time is that the norming groups to which the students are being compared are increasingly improving. This phenomenon may be especially apparent at the high school level because of the voluntary dropping out of lower ability students. The paper examined a technique called "reverse norming" in which norms are developed on only those students for whom test data is available for both points in time during which growth is to be measured. For example, if growth between the ninth and twelfth grades is to be examined, norms would be developed on only those ninth and twelfth graders who have test scores at both points in time.

The authors tried this technique to measure growth between the ninth and tenth grades. They compared the evaluation results using these norms to the results obtained when unrestricted local norms were used. Results showed no consistent differences. The authors suggested that the reverse norming be tried again over a longer time span.

Norming Issues: Cross-Sectional v. Longitudinal Norms
Equipercentile Assumption

Model A
I. Internal Validity
H. Interactions with Selection
1. Mortality

Baker, R., & Williams, T. Issues related to interpolation. NTS/TAC, 1978. Draft.

In order to calculate status at pretest or posttest time, when testing occurs more than two weeks away from norm dates, one needs to interpolate between norm dates in order to find standard scores corresponding to the testing date. The authors discuss several ways to interpolate and extrapolate and the extent to which they would yield the same results.

The first problem that one encounters is how to establish the norm date percentiles (or better, NCEs) to interpolate between. For example, suppose you find that when testing one and one-half months after the fall norm date that a class has a raw score of 22 corresponding to a standard score of 35. What NCE does this correspond to?

One way to do it is to look up the NCE equivalent of the standard score at pretest time (NCE_P) and the NCE equivalent at posttest time (NCE_S) and interpolate between them (see the right hand graph in Figure 1.1):

$$NCE_{\text{test time}} = NCE_P - (NCE_P - NCE_S)\left(\frac{\text{Date of testing} - \text{Norm date}_P}{\text{Norm date}_S - \text{Norm date}_P}\right)$$

In essence this procedure entails plotting equal standard score/raw score lines on a graph of NCEs versus dates.

Another way to calculate the NCE at test time is to plot equal NCE/percentile lines on a graph of standard scores versus dates. One would plot an observed standard score (SS_O) at the time of testing (see the left hand graph in Figure 1.1) and see which NCE lines it fell between. In order to arrive at the final result three interpolations would need to be calculated: (1) the standard score at the time of

testing corresponding to the upper (or high) NCE line (SS_H); (2) the standard score at time of testing corresponding to the lower NCE line (SS_L); and (3) the NCE line that passes through the SS_O and is parallel to NCE_H and NCE_L.

These calculations are:

(1) $$SS_H = SS_{H,P} + (SS_{H,S} - SS_{H,P}) \left(\frac{\text{Date of testing} - \text{Norm Date}_F}{\text{Norm Date}_S - \text{Norm Date}_F}\right)$$

(2) $$SS_L = SS_{L,P} + (SS_{L,S} - SS_{L,P}) \left(\frac{\text{Date of testing} - \text{Norm Date}_F}{\text{Norm Date}_S - \text{Norm Date}_F}\right)$$

(3) $$NCE_{\text{test time}} = NCE_L = (NCE_H - NCE_L) \left(\frac{SS_O - SS_L}{SS_{HS} - SS_L}\right)$$

where $SS_{H,P}$ = standard score for NCE_H at pretest time

$SS_{H,S}$ = standard score for NCE_H at posttest time

$SS_{L,F}$ = standard score for NCE_L at pretest time

$SS_{L,S}$ = standard score for NCE_L at posttest time

The authors feel that methods #1 and #2 will only yield the same results if there is a linear relationship between the standard scores and NCEs, i.e., that standard scores need to be normally distributed within the grade group at both norming dates. Given that smoothing and rounding occur when norms are developed, this may not always be exactly true. Therefore, the authors would recommend method #1.

One final issue dealt with is the proper norms tables to use for interpolation or extrapolation. The interpolation rule when not testing within two weeks of the norm dates of the test may be interpreted to mean that one should interpolate using the norms that bracket the test date, e.g., if a test has October and April norms and testing occurred in September, then the April norms for the previous year and the October norms for the current year should be used for interpolation. The authors

feel that this would be an unwise practice because it makes unproven assumptions concerning the growth rate over the summer months. The authors recommend rather that the current year's October and April norms be used to extrapolate backwards to the testing date.

Computation Problems:
Interpolation and Extrapolation

Model A
I. Internal Validity
D. Instrumentation

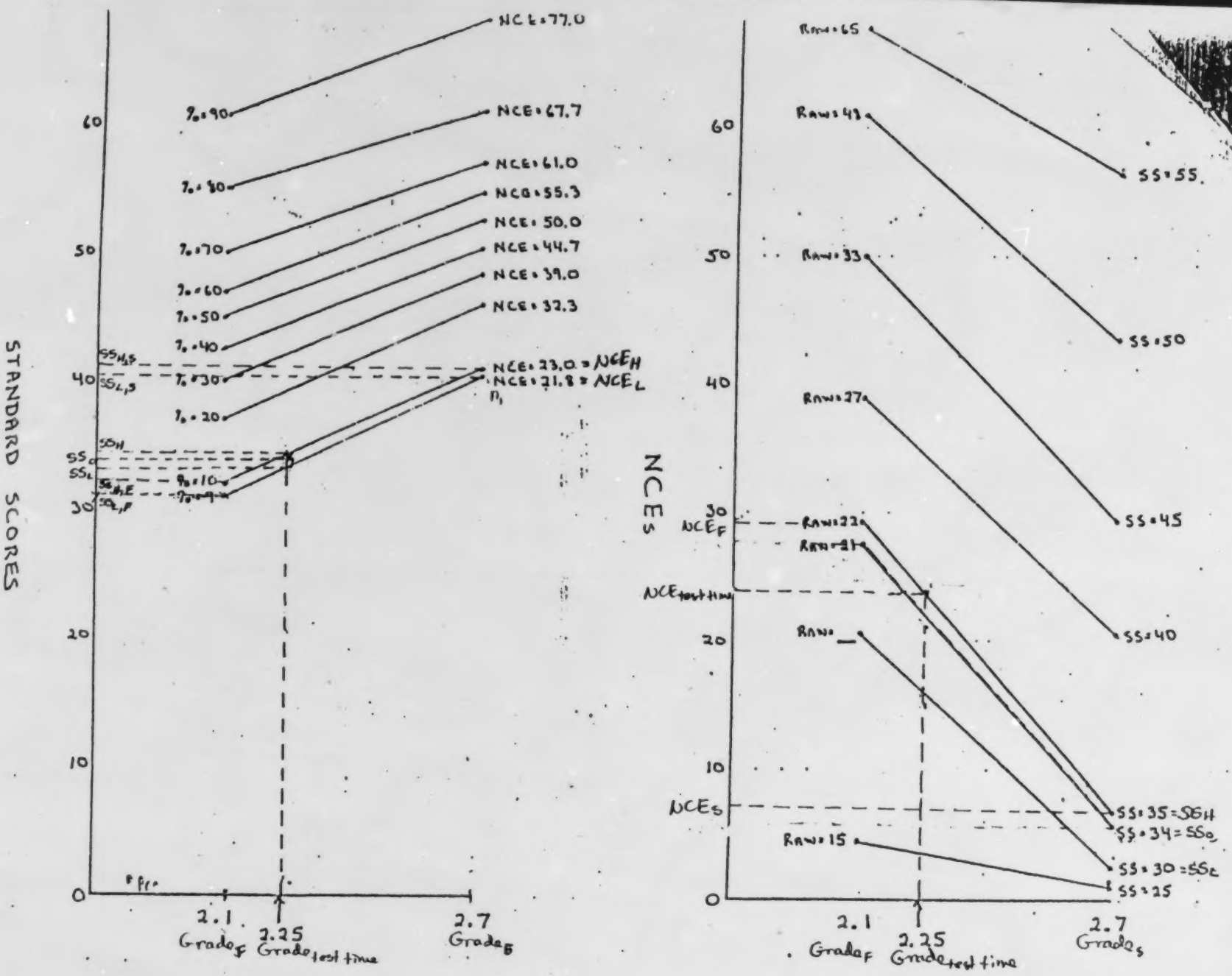
STANDARD SCORES
NCE = 77.0
NCE = 67.7
NCE = 61.0
NCE = 55.3
NCE = 50.0
NCE = 44.7
NCE = 39.0
NCE = 32.3
NCE = 23.0 = NCE_H
NCE = 21.8 = NCE_L
60
50
40
30
20
10
0
2.1
2.25
2.7
Grade_F
Grade_test time
Grade_S
NCE_S
Raw = 65
Raw = 43
Raw = 33
Raw = 27
Raw = 22
Raw = 21
Raw = 15
SS = 55
SS = 50
SS = 45
SS = 40
SS = 35 = SS_H
SS = 34 = SS_O
SS = 30 = SS_L
SS = 25
NCE_F
NCE_test time
NCE_S

FIGURE 1.1

Bridgeman, B. Extrapolation and interpolation in model A1 Title I Evaluation. ETS, 1978.

The author examines various interpretations of interpolation and extrapolation as it applies to projects not testing within two weeks of norming dates. Some interpretations are:

1. If a project actually began (or ended) before pretesting (or ended after posttesting), results might be extrapolated to the point where it began, rather than interpolating to the point where testing actually occurred. An example would be if the norms were developed for October 15, testing occurred on December 1 (one and one-half months later) and the program was fully operational by October 15. Month-per-month growth could be determined between the pretest and posttest, and one and one-half months of this growth could be subtracted from the December 1 pretest score in order to establish what performance would have been like if the students had been tested on October 15.

2. If a project actually began on December 1, then pretest percentile status would have to be interpolated from fall and spring norms. (The author notes that there is no easy way to do this.)

3. If pre- and posttesting occurred six weeks after fall and spring norming dates, respectively, then can one estimate posttest status by interpolating between spring of the current year and fall of the next year (i.e., over the summer months)? Or is the correct procedure to extrapolate using the fall and spring norms of the current year? The author demonstrates that the two methods can yield different results and recommends that extrapolation be used.

4. When year-to-year testing is used, scores are interpolated using the norms of the two years. This is done by calculating the proportion of the year that actual testing is away from norm dates and adjusting normative data by this amount. (This is the usual procedure and is analogous to getting an estimate of month-per-month growth.) In order to get the proportion, however, one must decide on the length of the school year--9, 10 or 12 months. This problem is related to the assumptions that one makes regarding summer growth. Results would be different depending on one's definition of summer growth.

The author concludes that "since the number of months between pre- and posttests in different projects could vary considerably, and assumptions used in extrapolations or interpolating could also differ, comparisons of the amount of gain among various projects must be made with extreme caution."

Computation Problems

Model A
I. Internal Validity
D. Instrumentation

Burton, B. Model A1 and the regression effect. Memo to TAC Staff. May 8, 1978.

Burton maintains that previous studies which argue that dual pretesting does not eliminate regression effects are not looking at a problem arising from the use of Model A1. A different model is being used (trivariate normal with all correlations less than unity) and therefore the results do not apply to Model A1.

Burton maintains that the equipercentile assumption implies that $r_{23} = 1$ (where 1 = testing at time #1, 2 = testing at time #2 and 3 = testing at time #3). This implies that $r_{12} = r_{13}$ which is the necessary condition for selection testing (time #1) to eliminate all regression effects.

Burton's paper seems to imply that the critical issue is whether the equipercentile assumption is correct. If it is correct, then separate selection testing eliminates all regression. If it is not correct then separate selection testing does not eliminate regression effects.

Regression
Equipercentile Assumption

Model A
I. Internal Validity
E. Regression

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPortland, J., Mood, A. M., Weinfeld, F. D. & York, R. L. Equality of educational opportunity. U. S. Department of Health, Education and Welfare, Office of Education. Washington, D. C.: Government Printing Office, 1966. (Also ERIC number ED 012 275.)

As part of an extensive survey of educational achievement and opportunities for American students, the authors compared the relative standing of various racial and demographic groups over time. The authors compared the achievement of the different groups in grades 6, 9 and 12 using scores on the Sequential Tests of Educational Progress that were standardized to have a mean of 50 and a standard deviation of 10. Using the white, metropolitan, Northeast subgroup as a comparison (this was the highest scoring group), the authors computed the number of standard deviation units that each group was below the reference group at each selected grade.

Results showed that on verbal ability and reading comprehension blacks and Puerto Ricans tended to fall increasingly behind the top group from grade 6 to grade 12 (most increases were from .1 to .5 standard deviation units). Results for other minorities are mixed. On the math tests, minorities tended to improve their relative positions, although blacks tended to increase only .1 standard deviation while other minorities increased .4 to .5 standard deviation units.

Comments:

1. Alternative explanation for results: cross-sectional rather than longitudinal norms.
2. Compared status of minorities to highest group of whites. May be more meaningful to compare them to the 50th percentile.

3. Results are probably not due to race per se but to SES.

Equipercentile Assumption

Model A
I. Internal Validity
H. Interactions with Selection

Committee on Use of Norms with Model A. Report. TAC Committee Meeting, Chicago, May 15-16, 1978 (A copy can be obtained from Joe Hansen, NWREL TAC, 710 S.W. 2nd Ave., Portland, OR 97204).

Two versions of the norm-referenced evaluation model were discussed by the committee. The first was the experimental approach. This approach, currently the one used by the TIERS, uses norms to generate a no-treatment expectation. That is, norms are used as a proxy control group.

Ideally, norms used for this purpose are based on students who are as similar as possible to the Title I group in question, who have not had the Title I treatment and who have received all other instructional treatments received by the Title I group except the Title I treatment. The committee decided that development of such norms is impossible. First, all known Title I students must be removed from the norms, as well as students in other compensatory programs. Any students at the same achievement level as Title I students who are left probably are not similar to Title I students in other respects. Historical norms are not adequate because of population or Title I program shifts. The committee concluded that Model A as currently defined does not yield a valid no-treatment expectation.

The other version of the norm-referenced evaluation model was a comparative approach. In this version, norms on Title I student gains would be developed each year. Individual student and program gains could then be compared to the national average each year. Further, comparison of the norms each year would provide a longitudinal data base on which to examine cumulative program success. Separate growth norms could be established for various periods of evaluation (fall to spring and spring

to spring). Although this system would allow for examination of longitudinal impact, it would not provide an estimate of absolute treatment effect.

Comments:

1. The solution discussed above only shows absolute growth and not growth relative to a no-treatment expectation. Why not just do pre-post t-test and report whether programs significantly increased achievement. It would save all the trouble of developing norms and would amount to the same thing.

Norming Issues: Contamination of Norms
Equipercentile Assumption
Alternative Models

Model A
I. Internal Validity
D. Instrumentation

Crowder, C. R. Relation of out-of-level testing to ceiling and floor effects on third and fifth grade students. Paper presented at AERA, Toronto, 1978.

This study investigated whether standard scale scores were comparable for in-level and out-of-level tests, and whether floor and ceiling effects with in-level tests were also evident for out-of-level tests.

Third grade students who had previously encountered floor or ceiling effects were tested both in-level and out-of-level on the Stanford Achievement Test. Fifth grade students were similarly tested with the Metropolitan Achievement Test.

The pattern of results that would be expected for students receiving an in-level and lower-level test is that, for students scoring at chance level on the higher level, the lower-level test would give lower average standard scores. For students scoring in the functional range of each test, the average standard scores should be the same, and students at the ceiling of the higher test should also be at the ceiling of the lower test.

The pattern of results that would be expected for students receiving an in-level and higher-level test is that students scoring at the floor in-level would also be at the floor on the higher level. Students in the functional range of both tests should average the same standard scores, and students who reached a ceiling on the in-level test should get higher standard scores on the higher test.

Results showed that these patterns were not consistent. The authors suggest that the differences in mean scale scores and percentages for the in-level and out-of-level tests may be due to the relatively large

increase in standard scores for raw scores at the extremes of the scale transformation.

Comments:

1. Alternative explanation for results: differences in test to curriculum matches between levels.

Norming Issues: Vertical Scaling

Model A
I. Internal Validity
D. Instrumentation

David, J. L., & Pelavin, S. H. Evaluating compensatory education: Over what period of time should achievement be measured? Journal of Educational Measurement, 15, 91-99, 1978.

(A longer technical report is Pelavin, S. H., & David, J. L. Evaluating long-term achievement: An analysis of longitudinal data from compensatory education programs. Stanford Research Institute, Menlo Park, California, EPRC 4537-15, 1977.)

The authors hold that, for policy making, the effectiveness of a compensatory education program must be evaluated in terms of student achievement for a greater period of time than the nine-month school year. A program should at least demonstrate sustaining effects through the summer following the program.

To determine whether the gains demonstrated during the regular school year are maintained over the summer, the authors conducted secondary analyses of longitudinal evaluation data from three separate evaluation studies. These studies included data on about 1500 disadvantaged students in grades 3-8. Different tests and metrics are used in the three reports of these studies.

Results showed that there can be large and statistically significant achievement losses over the summer both in terms of percentiles and grade equivalents. Fall to fall gains can be considerably smaller than fall to spring gains. The presence of this summer loss is enough to change conclusions about program effectiveness. The authors conclude that, given the policy implications of summer loss, all evaluations should be based on a 12 month evaluation cycle. They recommend that fall to fall evaluation be used so that the longer term effects of the program over the summer can be ascertained.

Comments:

1. Using a twelve-month cycle does not ensure that sustaining effects of the program can be shown. Results could also be due to lack of empirical norms and change in test levels.

Aggregation Problems: Fall-Spring v. Spring-Spring
Sustained Effects

Model A
I. Internal Validity
D. Instrumentation

Deck, Dennis. A Comparison of linear and normalized NCE transformation. Title I Technical Assistance Centers, Northwest Regional Educational Laboratory. Prepared for California Department of Education, January 1979. Draft.

Raw score data for third and sixth grade students on the CTBS (unit of analysis—some 1500 school means) were converted to NCEs two ways:

1. Normalized z-scores. Convert raw scores to national percentiles and then to z-scores based on the normal curve. Then, NCE = 50 + 21.06z.

2. Linear. NCE = $50 + 21.06 \frac{(\bar{X} - M)}{SD}$, where M and SD are the mean and standard deviation for the appropriate test and level of $\bar{X}$.

The two methods should be equivalent when X is distributed normally. The question is then, how normal are scores in school districts? The issue is of importance since in the TIERS conversions to NCEs, using norm-referenced tests is normalized z-scores, while that for nonnormed tests is linear.

Results showed that the use of linear rather than normalized z-score methods may result in moderate to substantive differences in mean NCE and in NCE gains. No consistent patterns were observed across tests, levels of tests or different administration times of the same test.

Linear v. Normalized
Conversion to NCEs

Models A, B, C
I. Internal Validity
D. Instrumentation

DeVito, P. J., and Long, J. V. The effects of spring-spring vs. fall-spring testing upon the evaluation of compensatory education programs. Paper presented at AERA, New York, 1977.

The authors compared the results of a fall-spring testing cycle with a fall-fall testing cycle using the reading subtest of the California Achievement Test for 238 disadvantaged students in grades 2-5. In some cases the test form as well as test level changed between testings; in other cases only the test form changed.

Results showed a significant decline in expanded standard scores between spring and fall in grades 2, 4 and 5. There was a decline in grade 3 but it was not significant. Large losses were also noted in percentile status. The authors claim that larger differences were observed when test levels were changed (grades 2 and 4) between the spring and fall testing period. This may be evidence of the problem of level equating.

The authors conclude that this is evidence for the summer loss phenomenon, and that aggregation should not occur between programs that are on fall-spring and spring-spring testing cycles.

Comments:

1. Alternative explanation for results: changes in test levels and forms.

Aggregation Problems: Fall-Spring v. Spring-Spring
Sustained Effects
Norming Issues: Vertical Scaling

Model A
I. Internal Validity
D. Instrumentation

Doherty, W. Restandardization study. System Development Corporation (undated).

Since publishers' norms tend to have less precision at the low end of the score distribution, and since an evaluation period may span two levels of a test, and because certain groups of students are totally unlike students in national norms, the present study constructed a new scale score using Levels 2 and 3 of the California Achievement Test with ESAA students in grades 3-5. Students in each grade were randomly assigned to take either Form A or Form B. All third graders took Level 2 and all fifth graders took Level 3; fourth graders were randomly divided so that half took Level 2 and half took Level 3.

Construction of the new scale score was performed using Thurstone's Absolute Scaling Method. Equating of the two forms was accomplished with the equipercentile method. These methods are described in this paper.

Results showed differences in raw score performance between Form A and Form B, which was more difficult. Differences in the shapes of the fourth grade distributions at the two test levels indicated large differences in difficulty across levels.

The author maintains that using the CTB scale scores for evaluating migrant programs gives nonvalid measures of growth, because the migrant population is so different from the national population. (This was demonstrated by looking at growth of these students over time.) The new scores are claimed to show more realistic growth patterns. The author suggests that the problems with the CTB scaled scores may also be present with other tests, and that level articulation problems will be encountered with any norms that have not been based specifically on disadvantaged, low-achieving students.

Norming Issues: Vertical Scaling
Sustained Effects
Equipercentile Assumption

Model A
I. Internal Validity
H. Interactions with Selection

Echternacht, G. The case for matching. ETS, Princeton, New Jersey, Memorandum for TAC Directors, May 11, 1978(a).

Echternacht discusses occasions which call for a comparison group to be designated through matching on pretest scores.

One argument against matching is that comparison students generally come from a population that has higher ability (because the lower-ability students have already been selected for Title I). Thus, there will be differential regression in the two groups.

Echternact suggests that one examine data from previous years in order to determine whether pre-post regression lines are parallel for the following groups: non-Title I students at Title I schools, and students at non-Title I schools who are matched with Title I and non-Title I students at Title I schools.

Parallel regression lines, Echternacht claims, are good evidence that differential regression is not likely. Echternacht argues that Model C is superior to Model B with such matching, but that if the matching requirements were fulfilled then Model B with matching would be superior to Model A.

Comments:

1. This solution controls for only one way in which groups can differ--different regression. Groups may vary systematically in other ways that would also result in poor comparisons.
2. After groups are matched would principal axis adjustment be used?

Model Selection/Bias
Regression

Model B
I. Internal Validity
H. Interactions with Selection

Echternacht, G. A note on selection and the norm referenced model. Draft. ETS Princeton, New Jersey, 1978(b).

Echternacht demonstrates that there is a way to generate a valid no-treatment expectation in Model A if the pretest is also used to select students. Basically the result is

$$y_o = \mu_y - r_{xy} (\sigma_y/\sigma_x) (x - \mu_x)$$

where y_o is the expected posttest score, μ_x, σ_x are population mean and standard deviation on the pretest, μ_y, σ_y are population mean and standard deviation on the posttest, x is a pretest score, and r_{xy} is the correlation between pretest and posttest scores.

Echternacht has several cautions for use of these results. (1) Many of the variables are population statistics obtained from norm group statistics. Most norms are cross-sectional which means that some statistics, such as r_{xy} may not be available. (2) Norm group means may not provide good means for use in the formula because they may include some project students.

Echternacht maintains that it would be impossible to obtain good estimates of project effects using his formula unless a district could generate local longitudinal norms using students not in Title I. If such norms are not available or cannot be developed, then the pretest should not be used for selection.

Regression

Model A
I. Internal Validity
E. Regression

ESEA Title I annual evaluation report: FY 1975. Office of Federal Programs, South Carolina Department of Education, Columbia, South Carolina, November, 1975.

In the South Carolina evaluation report Model A results are reported separately for students who were evaluated spring-spring and for those who were evaluated fall-spring. Data is reported for over 24,000 students in grades 2-12; about 40% of the students were tested spring-spring and the remaining 60% were tested fall-spring. CTBS and CAT reading and math tests were given.

Results were that in 32 cases where comparisons could be made, the differences between actual posttest scores and expected posttest scores of fall-spring were, in all but two cases, much larger in a positive direction than corresponding differences for spring-spring testing.

One possible weakness of these results was that different sets of students were tested fall-spring and spring-spring. It is possible that differences between the student groups resulted in the differences, rather than the testing cycle itself. However, with such a large number of students, systematic selection or program bias may be less likely.

Comments:

1. Alternative explanation for results: used interpolated rather than empirical norms. Thus, fall or spring performance may be incorrectly estimated.

Sustained Effects

Model A
I. Internal Validity
D. Instrumentation

Estes, G. D., & Anderson, J. I. Observed treatment effects with special regression evaluation models in groups with no treatment. Paper presented at the annual meeting of the American Educational Research Association, Toronto, March 1978.

Estes and Anderson applied the Special Regression Model in a situation where no treatment was provided to what would have been defined as the Title I group under the TIERS guidelines. Samples for the analyses were drawn from 730 ninth grade students enrolled in mathematics classes in schools "somewhat similar" to the Title I schools in a large city school district. Three different achievement tests in mathematics were used to define three nonindependent samples for the analyses. "Cutoff scores" comparable to those used to assign students to ESEA Title I math classes were established for each test. Each test was then identified with its own "treatment group" and "comparison group."

Analyses provided differences between the "treatment group" and the "comparison group" at the "cutoff score" and at the pretest mean of the "treatment group." Tests of statistical significance were applied to each estimate of treatment impact. All but one of the "t" tests were significant at the .05 level. Two estimates indicated negative effects and three indicated positive effects. Thus, treatment effects were found where there were no treatments.

The authors suggest that the presence of even minor floor and ceiling effects, which violate an assumption of the Special Regression model, might have a significant effect on assessing project impact. They advise evaluators to attend closely to model assumptions. They suggest evaluators might delete scores of students who contribute to floor effects (in the treatment group) and ceiling effects (in the comparison group).

It is recommended that within group correlations used in the Special Regression Model be interpreted after the effects of restriction in range are taken into account. This will be particularly important in comparing the level of obtained correlations against the guidelines for the Special Regression Model.

Finally, the authors recommend that additional guidelines are necessary to assist in interpreting results from the Special Regression Model. They suggest that effects of other projects as well as some absolute standards may be useful aids to interpreting results.

Model Selection/Bias
Floor/ceiling Effects

Model C
I. Internal validity
D. Instrumentation

Faddis, B. J., & Estes, G. D. Fall-to-spring vs. fall-to-fall evaluations of a large Title I program with a comparison group design. Paper presented at AERA, Toronto, 1978.

The authors feel that differences in evaluation results between fall-fall and fall-spring evaluation periods would only be a problem for Model A unless there is differential summer loss for the treatment and comparison groups. To determine whether the evaluation period affects results using a Model B design, the authors compared treatment effects in NCEs for a fall-fall and a fall-spring evaluation cycle for students on the reading subtest and students on the math subtest of the CTBS. The authors found no treatment effects for either project on either evaluation cycle, and no differences between results depending on cycle.

The authors state that these results are only suggestive because of problems with large attrition rates, no duplication of the summer loss phenomenon found in other studies and because the CTBS had no empirical fall norms (hence fall estimates of status may be inaccurate).

One subresult from this study was the determination of percentile status over two years for a Model C and a Model B comparison group. Students comparable to Title I students, but not served, lost 1.4 percentiles in reading and gained 3.9 percentiles in math from one year to the next. The Model C comparison group gained 3.8 percentiles in reading and 1.5 in math. Thus, neither group maintained its relative position over two years.

Sustained Effects
Equipercentile Assumption

Models A, B, C
I. Internal Validity
D. Instrumentation

Model B
I. Internal Validity
H. Interactions with Selection

Fishbein, R. L. The use of non-normed tests in the ESEA Title I evaluation and reporting system--some technical and policy issues. Paper presented at AERA, Toronto, 1978.

This paper examines several technical and policy issues related to the use of nonnormed tests in the ESEA Title I Evaluation and Reporting System. The first issue discussed was the technique of converting program results into NCEs when a CRT is used. In order to do this, one must estimate the standard deviation of the national population on the CRT using the ratio of the local group's standard deviation on the CRT and the nationally normed test. Fishbein states that thus far there has been no information presented on this topic.

The second issue presented by Fishbein concerns the use of a linear transformation in Models B2 and C2 to translate the gain on CRTs into NCEs. The typical method is through normalized z scores. The two methods would only be equivalent if the underlying score distribution for the test was normal. In most CRTs, Fishbein claims, pretests are designed to be positively skewed and posttests to be negatively skewed. Therefore, conversions by the two methods would not be the same. This would argue against the feasibility of aggregating results for CRTs and NRTs.

Floor and ceiling effects may also cause problems in Model A2. In order to generate a no-treatment expectation in Model A2 it is necessary to equate a criterion-referenced test to a norm-referenced test through use of equal percentile equating. If distributions on the CRT are very skewed at pretest and posttest times this may be impossible to do because there may be no cumulative frequency on the pretest CRT that corresponds to the median raw score value on the posttest CRT.

Fishbein then discusses a seeming paradox that occurs when using CRTs. On the one hand, in order to make a testing program using a CRT manageable, the number of objectives that are tested must be fairly small. However, if the number of objectives on the CRT is small then one might not get a good measure of overall improvement in reading which is the major goal of compensatory education. For example, Fishbein says "if the objectives of a program are narrowly defined, one could develop instruments to measure mastery of those objectives and perform a technically correct evaluation using Model A2 or C2. If the gain is large, this means that the students have shown more growth on the narrow set of objectives than would have been expected without the special program. But what about the possibility that because so much instructional time has been devoted to meeting the narrow goals of the program, other important curriculum areas have been neglected?" Fishbein suggests that perhaps a series of narrow-objective CRTs and norm-referenced tests be given in order to test both sets of objectives.

Nonnormed Test Problems:
- Estimating SD of National Population
- Linear v. Area Transformation to NCEs
- Floor and Ceiling
- Local v. National Objectives

Models A, B, C
- I. Internal Validity
  - D. Instrumentation
- V. Construct Validity of Effect
  - B. Mono-Operational Bias

Model B
- I. Internal Validity
  - D. Instrumentation

Fuentes, E. J., & Jeffress, E. L. The use of teacher estimates of children's achievement in selecting appropriate test levels vs. blanket out-of-level testing: The problem of an efficient method of matching the student's functional level with the proper test level. Center for Educational Research and Evaluation, Research Triangle Institute, Research Triangle Park, North Carolina 27709. Paper presented at AERA, Toronto, 1978.

This paper reports on a feasibility study concerning the relative merits of blanket out-of-level testing versus test level assignments based on teacher estimates of each child's achievement level. The study involved ESEA Title I Migrant students from ten schools in grades 2, 3, 4 and 5. Each child was given the CTBS reading and math subtests one level below his/her grade placement. Teachers also estimated each child's achievement level in reading and math.

The percent of students in each grade scoring at chance levels on the out-of-level tests ranged from 5%-26% for reading and 1%-22% for math. As grade level increases, the difference between grade equivalent also increased. The authors concluded that blanket out-of-level testing resulted in an acceptable match of children's achievement levels and test difficulty.

To assess teachers' ability to match children and test levels, teacher estimates of achievements were correlated with obtained CTBS grade equivalent scores in reading and math. Teacher estimates tended to be lower than obtained test scores, with the average correlation being .70 in reading and .62 in math. The authors concluded that teacher predictions were not accurate enough to justify individual level selection as opposed to blanket out-of-level testing.

Norming Issues:
Vertical Scaling
Procedures for Out-of-Level Testing

Model A
I. Internal Validity
D. Instrumentation

Glass, G. Regression effect. Memo to Unknown, Interested Parties, March 8, 1978.

This paper contains a mathematical proof and an explanation of why dual pretesting does not eliminate regression effects. Glass begins with a proposition that is necessarily implied if dual testing does eliminate regression, if there is no treatment effect, (namely that $\bar{X}_1 < \bar{X}_2$, and $\bar{X}_2 = \bar{X}_3$, where subscripts are 1 = pretest #1 (selection), 2 = pretest #2 and 3 = posttest), and shows that this leads to an impossible conclusion.

First, an "article of unshakable faith" is that the regression effect is proportional to the correlation of the two variables involved, i.e.,

$$\text{Regression from \#1 to \#2} = r_{12}\left(\frac{\bar{X}_1 - \bar{X}}{SD_x}\right) - \bar{X}_1$$

where X is the mean of the entire group from which the subgroup was selected.

Also,

$$\text{Regression from \#1 to \#3} = r_{13}\left(\frac{\bar{X}_1 - \bar{X}}{SD_x}\right) - \bar{X}_1$$

Since $\bar{X}_2 = \bar{X}_3$ (and, therefore, regression from time 2 to time 3 equals zero), r_{12} must equal r_{13}. But this is false since all psychometric measures correlate more highly the closer together they are in time.

Glass then provides a logical explanation of why r_{12} would not be expected to equal r_{13}. "The misconception stems from the belief that 'measurement error' is an intrinsic property of a test instead of a characteristic of a particular use of a test. There are different measurement errors, with different variances, depending on whether one is

speaking of one-day reliability or one-year reliability. So although the double-pretest design takes care of the 'errors of measurement' between pretest #1 and pretest #2 (as such errors exist over similar intervals of time), it does nothing about the 'errors of measurement' that exist for the interval between pretest #2 and posttest.

"Notice that if the 'errors of measurement' are intrinsic to a test, then its reliability would be the same whether assessed as stability across one hour, one week or one year. Hence, r_{12} would equal r_{13}..."

Regression

Model A
I. Internal Validity
E. Regression

Goldberger, A. S. Selection bias in evaluating treatment effects: Some formal illustrations. Institute for Research on Poverty, Discussion Papers, University of Wisconsin--Madison, April 1972.

Goldberger discusses a statistical artifact that might tend to bias estimates of treatment effect in Model B using an analysis of covariance procedure for estimating the no-treatment expectation. Goldberger examines a previous observation that "a spurious, deleterious effect of the treatment will be observed when the selection procedure systematically puts lower-ability subjects into the treatment group and higher-ability students into the control group." Goldberger found that the spurious effect arises when subjects are selected for treatment explicitly on the basis of true ability (the assumption made in earlier studies). However, when selection is based explicitly on the basis of pretest score, the spurious effect vanishes.

In his examination, Goldberger considered the case where the treatment effect is zero. In this case, when posttest is regressed on pretest with a dummy variable for treatment, and selection is based on pretest score, the dummy variable receives zero weight because all the information in the dummy variable is also contained in the pretest score. When selection is on the basis of true score, however, the dummy variable (i.e., group membership) does not align exactly with pretest scores. Thus, there is some information in the dummy variable that is not contained in the pretest scores, and the regression weight on the dummy treatment variable will be nonzero.

Goldberger uses a rigorous mathematical proof to support his argument. The implication is that Model C, which does require explicit selection on the pretest, does not fall victim to the spurious effect

caused by selection on true ability. However, the author points out that the standard error of treatment effects generated by Model C is considerbly larger than standard errors for Model B.

Model Selection/Bias

Haenn, J. F., and Proctor, D. C. A practitioner's guide to out-of-level testing. Paper presented at AERA, Toronto, 1978.

As the title states, this paper is a practitioner's guide to out-of-level testing. The paper begins with a definition of out-of-level testing, reviews the pertinent literature and discusses the reasons for testing out of level. Guidelines for selecting the appropriate test level are presented and procedures for making score conversions are outlined. Three appendices are included which give chance level scores and their corresponding percentiles and grade equivalents for commonly used achievement tests, names of expanded standard scores for these tests and worksheets for converting to in-level scores.

One interesting result from this wealth of data is that chance-level scores can correspond to percentiles 2 to 39 depending on the test and level.

Procedures for Out-of-Level Testing

Model A
I. Internal Validity
D. Instrumentation

Hardy, Roy. A comparison of model A and model C: Results of first year implementations in Florida. ETS, Evanston, IL, personal communication, January 1979.

The author compared NCE gains from Models A and C using data from Florida school districts. Gains for Models A and C were aggregated across grades and sites and the results compared. Model A showed results that were consistently greater than those in Model C.

The author notes the following limitations to the study: (1) It may be that the sites that chose to use Model A were systematically different from those that chose Model C. The author examined this and found that similar sized school districts were included in both samples and that pretest averages were very close. (2) Different testing intervals were used for Models A and C. Most Model A projects tested fall-to-spring while most Model C projects tested spring-to-spring. (3) Gains were aggregated across grade levels. The balance of grades represented in each may not be the same.

The author also reports on a direct comparison of Models A and C that was done in a single school district. Results showed a larger gain for Model A. But, again, Model A used a fall-to-spring design while Model C used a spring-to-spring design.

Model Selection/Bias

Model A
I. Internal Validity
D. Instrumentation

Hiscox, S. B., & Owens, T. R. Behind the basic assumption of model A. Paper presented at AERA, Toronto, 1978.

The authors examined the validity of the equipercentile assumption in Model A. The authors first discuss various factors that might introduce unknown amounts of bias into the estimates made on the basis of the equipercentile assumption. (1) If a norm group is used which is not similar in many respects to Title I students, estimates may be off because growth curves of different populations cannot necessarily be assumed to be the same. (2) As students drop out of high school, the score needed to achieve a given percentile level from one testing time to another will increase. This is because presumably the lower students are dropping out and the norm group is getting increasingly better. (3) Most test publishers use cross-sectional rather than longitudinal norms. Equipercentile estimates might not be the same for the two because of population shifts, curriculum shifts, etc. (4) There are varying amounts of error in articulating forms and levels of tests. Thus, using one form or level of a test to predict performance on another form or level may introduce an unknown amount of bias. (5) If a test is used to measure progress which does not adequately sample the curriculum of a program, then students may not show enough achievement on the test to maintain their percentile rank. (6) Although Model A substantially controls for one type of regression by requiring that students be selected for the program on a basis other than the pretest, there is another type of regression that may influence estimates. That is, "given a population measured on a specific variable and a sample selected from the population, the best prediction of the sample's mean measurement at a

second point in time is not the sample mean but, rather, the sample mean (in standard deviation units) multiplied by the test-retest reliability of the measurement instrument."

The authors attempted to examine the equipercentile assumption by tracing the percentile standing of Title I and non-Title I students over four years. Four hundred nineteen students were followed from grades 4-7 and 327 students were followed from grades 7-10. To determine whether percentile standings vary depending on initial achievement standing, groups were divided up by percentile standing (at or below 5th, 15th, 25th and 35th).

Results showed that while many of the groups showed fairly consistent percentiles, others showed enough change over the last three years to cause problems with an evaluation based on an assumption of percentile maintenance. About the same amount of variability was noted in the Title I and the ron-Title I students. It is noted by the authors that students eligible for Title I over longer periods of time do not show consistent increases in their percentile levels. The authors stated that it is difficult to tell whether the variability is due to real differences in student achievement, changes in test levels administered over the four years or the amount of incomplete data involved in the study.

The final analysis conducted by the authors was to average scores for all students in each grade for each number of years of Title I eligibility. Results showed very little variability in percentile rank from year to year for both Title I and non-Title I students (most differences were between 2 and 3 percentiles). The authors conclude by

saying that there is little evidence in these data to suggest that either Title I students or comparable low-scoring non-Title I students might show percentile declines.

Equipercentile Assumption	Model A
Norming Issues:	I. Internal Validity
Cross-sectional v. longitudinal	D. Instrumentation
	E. Regression
Scaling	H. Interactions with Selection
Test/Curriculum Match	Model B
Regression	I. Internal Validity
	F. Selection

Holthouse, N. D., Stofflet, F. P., & Tokar, E. B. Achievement, social class and the summer vacation. Norfolk Public Schools, Norfolk, Virginia, 1976.

The authors attempted to determine whether regular students tended to lose or gain in test scores over the summer. They chose 2666 students from grades 1-10. All students in grades 1-6 took the SRA in-level and all students in grades 7-10 took the STEP in-level. Results showed significant gains in reading, math and language arts normalized standard scores over the summer. This gain was comparable for all social class and income levels. However, the authors found that blacks made significant gains in all three subject areas while whites tended to not gain or lose over the summer.

Comments:

1. The significant summer gains described above are based on standard score differences of only 1/20 to 1/7 of a fall SD. Therefore, while significant, gains may not be educationally meaningful.
2. Implications of this study: (a) Test publishers are correct in building summer growth into the norms. (b) The equipercentile assumption may be incorrect for some groups.

Equipercentile Assumption
Estimate of Summer Growth

Model A
I. Internal Validity
D. Instrumentation

Model B
I. Internal Validity
F. Selection

Jaeger, R. M. On combining achievement test data through the NCE scaled score. University of North Carolina at Greensboro, undated.

This paper explores the issue of aggregating Normal Curve Equivalent (NCE) data across tests. Jaeger reanalyzed data from the Anchor Test Study (ATS). He concludes that aggregating across tests is unwarranted. The lines of evidence are several.

First, Jaeger shows that there is sampling error in national norms by comparing the norms derived in the ATS study to publisher's norms. In other words, percentiles presented by the major tests differ randomly from those that would be developed from the true national population. Thus, percentiles used in NCE conversions do contain some error. The amount of error depends on the test. One result of this sampling error is that a class taking two tests would probably have different percentile statuses on each simply because of the error in the norms. This is in addition to any differences that may arise because one test may measure a curriculum better than another test. Jaeger acknowledges that it is a matter of opinion whether the amount of error is enough to worry about. There has been some debate on this point.

Another reason why data should not be aggregated across tests, in Jaeger's opinion, is that effects as measured by various tests are different. Jaeger chose a certain percentile effect (say from the fifth to the tenth percentile) and determined the raw scores necessary to achieve this effect on various tests, using the ATS norms. Then, he used the original publisher's norms to find the percentile effect represented by the raw score effect. He found that the effects varied from the original percentile effects and varied between tests. He concludes that

since a group would show different effects depending on the tests they took (because of error in the national norms), results should not be aggregated across tests. The practical implications of these findings are controversial. It is unclear whether the differences discovered are significant.

Third, Jaeger maintains that one cannot aggregate data between NRTs and CRTs. Calculations on the Title I Evaluation and Reporting System reporting forms use a linear conversion to transform data from CRTs to NCEs, while the conversion of the results from NRTs is based on an area transformation. The two techniques will be equal only if the underlying score distribution is normal. Jaeger maintains that this is not likely to occur when using CRTs because they are frequently designed to have skewed distributions. Differences between the two methods (pretending that a NRT is a CRT and calculating NCEs in the two methods) were as much as three NCEs. Jaeger maintains that the NCEs generated by the two methods are large enough to make a difference.

Other issues which Jaeger touches on are: (1) In order to use the models with CRTs one must equate the CRT to an NRT. To accomplish this, two tests must be highly correlated. He maintains that there is evidence that such correlations are often small. (2) One should not depend on NCEs for vertical scaling. There are problems with vertical scaling. NCE equating would only add to these problems. (3) Although tests reported in the ATS were reading tests, Jaeger believes that there would be more equating error in math tests.

Comments:

1. There is much controversy concerning all the points made by Jaeger.

Norming Issues:
  Sampling Error
  Vertical Scaling
Test/Curriculum Match
Nonnormed Test Problems:
  Linear v. Area Transformation to NCEs
  Equating Nonnormed Test to NRT
Aggregation Problems:
  Across Tests
Problems with NCEs

Models A, B, C
I. Internal Validity
   D. Instrumentation

Model A
I. Internal Validity
   D. Instrumentation

Jenkins, J. R., and Pany, D. Standardized achievement tests: How useful for special education? Exceptional Children, 1978 44, 448-453,.

The authors matched up basic reading curricula with various well-known standardized achievement tests to investigate the hypothesis that various tests measure student achievement differently. That is, students in a particular curriculum taking two different tests would receive different grade equivalent scores on the two tests. Bias in test results was estimated by comparing the relative overlap in the content of four separate reading achievement tests with the contents of five commercial reading series at first and second grade levels. Overlap between each achievement test and each reading series is reported in terms of achievement test grade equivalent scores that would be expected given that a student got correct just those words on an achievement test that appeared in a reading series.

Results indicate clear discrepancies between the grade equivalent scores obtained--both between tests for a single curriculum and on a single test for different reading curricula. The authors concluded that for young children, achievement test scores are not useful for program placement, program planning or program evaluation because the choice of a test affects the status and gains of students and classes.

Comments:

1. These results assume, that students learn everything that they are exposed to in school and that responses to test items are completely determined by the content of the the reading curriculum. If either of these assumptions do not hold, then differences between tests may

not be as great. Also, these results may argue as much for selecting tests that match your curriculum as it does for not using standardized achievement tests.

Test/Curriculum Match

Models A, B, C
I. Internal Validity
D. Instrumentation

Kaskowitz, D. H., & Norwood, C. R. A study of the norm-referenced procedure for evaluating project effectiveness as applied in the evaluation of project information packages. Stanford Research Institute Research Memorandum URU-3556, Melo Park, CA, January 1977.

SRI used the norm-referenced model to evaluate PIPs. Part of this evaluation was a thorough examination of the pitfalls of the model.

The first potential weakness discussed had to do with assumptions underlying the use of standardized tests. All persons seem to agree that the content of a test used to evaluate a project should match the content of instruction. Persons who advocate the use of CRTs claim that they have fewer irrelevancies. Those who advocate the use of NRTs believe that a program is effective only if it shows general gains in achievement above those specifically claimed by the program.

Besides the policy assumptions described above, a second set of assumptions relate to the validity of the norms. The soundness of a norm-referenced evaluation depends on the extent to which assumptions used when designing the norms are valid. Possible problems here are cross-sectional rather than longitudinal norms, procedures used in vertical scaling and assumptions concerning summer growth built into the norms. Any of these items could introduce spurious effects into estimates of normal growth. The authors examined these issues for the Metropolitan Achievement Test (MAT) and found the possibility for some spurious effects.

The third possible weakness of the norm-referenced model is the possibility that the equal percentile assumption is incorrect for the types of students served in Title I programs. These students may lose ground over time, and hence, the no-treatment expectation generated by

the equal percentile assumption may be too great. The authors examined the validity of the assumption by comparing equal percentile growth curves (the curve generated by plotting pretest standard scores against the posttest standard score predicted by the equal percentile assumption) to actual growth curves of students (the curve generated by plotting pretest standard scores against the average observed posttest score for all students having a given pretest score). Further, the authors calculated summary statistics on whole groups in order to determine whether students drop in percentiles over time.

Results indicated that (1) A non-follow-through group tested three years in a row lost ground with respect to the norm group, i.e., their percentiles decreased. (2) A second non-compensatory education group with low socioeconomic status showed gains in percentiles over the course of fall to spring. The authors feel, however, that this group is not representative of Title I students. Also, results are not based on a full year, but only fall to spring. (3) Confounding these results may be the effects of misarticulation between test levels and testing that occurred off of norm dates in different directions. (4) The empirical growth curve for the non-compensatory education group tracked the equipercentile curve very closely in the midrange of pretest scores, but tended to lie above the equipercentile curve for extremely low pretest scores. (5) Other comparisons between empirical and equipercentile growth curves are nonsystemic between groups, subgroups and grades. This editor feels that no clear pattern emerges. The authors of the paper concluded that: (a) For pupils with extremely low pretest scores, the equipercentile assumption leads to a predicted posttest standard score

that is much lower than what was observed in the MAT longitudinal data, the non-follow-through student data and the compensatory reading student data. (b) Use of the norms based on the standardization group will lead to an expected posttest score that will be too high for students ordinarily in compensatory programs, especially minority students who have pretest scores that are not extremely low. Thus, results depend on beginning status of the students and other characteristics.

A fourth possible weakness of the norm-referenced evaluation is that the gains in standard score necessary for normal growth may vary considerably over grade level (and over pretest scores within grade levels) under the equipercentile assumption. That is, it may be harder to achieve normal growth (and to achieve more than normal growth) as grade level increases. To demonstrate this, the authors compared the gains necessary for normal growth on the MAT to those additional gains necessary for educationally significant growth (the 1/3 standard deviation rule). Since gains necessary for normal growth decrease with grade and standard deviations of standard scores tend to increase with grade, the 1/3 standard deviation criterion is an increasing multiple of normal growth. Therefore, while at the lower grades educationally significant growth may be one to one and one-half times normal growth, at the upper grades, it may be five times normal growth. Thus, educationally significant growth is harder to attain at the upper grades.

The fifth possible weakness of the norm-referenced model discussed by the authors relates to computational procedures for statistics. One issue is whether the average of the posttest standard scores as predicted by the equipercentile assumption is the same as the predicted posttest

standard score of the average. This problem relates to the unit of analysis and will be the same if the function relating standard scores to percentiles is linear. Linearity depends not only on the equations (if any) which are used to generate the norms tables, but also on conventions for norms tables use, such as how to calculate a nontabled value. For the MAT the authors found that the "equipercentile function" is linear. However, if the equipercentile assumption is not appropriate for disadvantaged students, then it must be shown that the alternative function is also linear.

In general, the authors conclude that: (1) Standardized tests should only be used selectively; that is, if test content corresponds to program or policy objectives. (2) Cross-sectional norms are not adequate for predicting longitudinal gains. (3) The inadequacy of the equipercentile assumption for disadvantaged populations tended to be confirmed, although modifying the assumption tended not to change conclusions about the program. This may be a problem in evaluations which extend over more than one school year. (4) The educationally significant growth criterion (1/3 standard deviation) is not uniformly stringent across grade levels. (5) Changing the unit of analysis does not appear to change conclusions regarding program impact.

Test/Curriculum Match
Local v. National Objectives
Norming Issues:
  Cross-sectional v.
    Longitudinal Norms
  Vertical Scaling
  Summer Growth
Equipercentile Assumption
Computation Problems:
  Unit of Analysis

Models A, B, C
I. Internal Validity
    D. Instrumentation
II. Statistical Conclusion Validity
    A. Power

Model A
I. Internal Validity
    D. Instrumentation
    H. Interactions with Selection

Aggregation Problems:
Across Grades

Model B
I. Internal Validity
F. Selection

Linn, R. L. Evaluation of Title I via the RMC models: A critical review. Paper presented at the CSE Invitational Conference on Measurement and Methodology, Center for the Study of Evaluation, UCLA, Los Angeles, January 4-5, 1978.

Linn provides a general critique of the Title I Evaluation and Reporting System. Then he deals more specifically with the Norm-Referenced Model and the assumptions on which it rests.

To begin, Linn describes four strengths which he feels deserve special mention. They are:

1. The recommendation that the same test level be given pre and post.
2. The requirement that normed tests be given near the norming dates.
3. The use of NCEs rather than the more familiar grade equivalents as a common reporting metric.
4. The sound logical and empirical basis for Models B and C (however, Linn feels the assumptions will rarely be met in practice).

After a brief description of Models A, B and C, including the provision for using local norms and criterion-referenced tests, Linn raises three limitations concerning the TIERS. First, no attention is given to the problem of treatment definition. Second, achievement in reading and math as measured by standardized, norm-referenced tests, is accepted as a sufficient outcome. Third, although NCEs are an improvement over grade equivalents, they are subject to the same strengths and weaknesses of other normalized standard scores. Most importantly, NCEs are dependent on norms data and they require strong

assumptions to support the equal interval claim made for them. Linn acknowledges that while the issues surrounding these limitations are outside the confines of what the RMC models are purported to do, these issues nonetheless call for more justification of the choices made.

Linn then examines problems of estimating a no-treatment expectation in experimental and quasi-experimental research. He follows this by commenting on the manner in which RMC dealt with these problems in defining the Title I Evaluation and Reporting System, especially the Norm-Referenced Model. Linn stresses that in the absence of random assignment and controls over treatment and control groups proper adjustments (i.e., estimations) are unknown and can only be approximated with unknown accuracy.

It is from this point of view that Linn examines the basis for the equipercentile growth assumption, the estimator procedure for the Norm-Referenced Model and the principal-axis or standardized-gain-score adjustment method used in the Comparison Group Model. Linn argues that these estimation techniques are consistent with one another and that both depend upon the validity of the so-called "fan spread hypothesis." This hypothesis states that in the absence of treatment effects groups will maintain their same average percentile rank or standard score position within the distribution of individual scores. When expanded scores such as grade equivalents are used longitudinal growth curves show groups "fanning out" from population averages.

Estimation procedures based on the fan spread hypothesis require a rather strong assumption which, if unsound, can lead to biased estimates. Linn addresses the fan spread hypothesis by examining data

from two studies. These data raise doubts about the generalized soundness of the fan spread hypothesis. He concludes that the key assumption of Model A is tenuous.

Next, Linn addresses the soundness of treating NCE results as equivalent regardless of the test. He cites examples drawn from Jaeger's analysis of the equivalence question which used Anchor Test Study data. He concludes that choice of a test (even excluding content appropriateness as a factor) can lead to a biasing effect of three to four NCEs. Whether this level of bias can be tolerated depends, according to Linn, on the likely size of project effects as registered on standardized, norm-referenced tests.

Finally, Linn examines the extra error that may be introduced in Model A2 due to equating a non-normed test to a normed test. In order to adequately equate tests, they should measure the same thing, i.e., have a high correlation. Linn showed that even tests that correlate highly might result in equating errors that could change the observed NCE gain. He took two normed tests and pretended that one was a non-normed test. He calculated the no-treatment expectation for a series of pretest scores for this "non-normed" test in two ways--by equating the test to another normed test, and by using the actual norms of the test. Differences ranged from 1/3 to 1/2 of a norm group standard deviation.

Linn concludes his critique by emphasizing the tension between technical soundness and feasibility that the Title I Evaluation and Reporting System plunges into. Model B rates high on soundness but low on feasibility. Model A, on the other hand, rates low on soundness but high on feasibility. As a result, while Linn believes that the Title I

Evaluation and Reporting System would lead to improvement in practice, he judges that the System will fall short of providing valid, comparable and interpretable results.

In a broader sense Linn feels that "the enterprise of educational evaluation is too immature to be ready for a fixation on a small set of models." He suggests that we may [illegible] by moving away from the notion that each year's evaluati[illegible] stand on its own. Systematic use of historical data may provide a sounder basis for comparison than "control" groups or national norms.

Treatment Definition
Problems with NCEs
Equipercentile Assumption
Nonnormed Test Problems:
  Equating Nonnormed Test
    to NRT
Aggregation Problems:
  Across Tests

Models A, B, C
IV. Construct Validity of Cause
V. Construct Validity of Effect
    B. Mono-Operational Bias

Model A:
I. Internal Validity
    D. Instrumentation
    H. Interactions with Selection

Model B
I. Internal Validity
    F. Selection

Long, J., Horwitz, S., & DeVito, P. An empirical investigation of the ESEA Title I evaluation system's proposed variance estimation procedure for use with criterion-referenced tests. Paper presented at AERA, Toronto, 1978.

This paper examines the tenability of the Title I Evaluation and Reporting System procedure for estimating the national group standard deviation on a local criterion-referenced test using the local group's standard deviation on the normed and non-normed test. This procedure assumes that the ratio of the local to the national standard deviation will be the same for the normed test as for the non-normed test.

The authors used test scores on 318 students in grades 2, 3, 4, 7 and 8. Standard deviations for the total reading score on the CTBS were used to estimate the population standard deviations for the CAT. Results showed that at all grade levels studies proceeding under the assumption that the ratio of local to national standard deviations would be the same for the normed as for the non-normed test led to an estimated standard deviation of the national norm group which was higher than the actual standard deviation for the national norm group obtained by inspection from the publisher's technical manual. The authors also note that had the CAT been chosen as the norm-referenced test and the CTBS chosen as the criterion-referenced test, the opposite results would have been observed--that is, the population standard deviations would have been systematically underestimated.

The authors conclude that "in any estimation procedure there will, of course, be errors of estimation. The degree to which those errors are systematic or random should be investigated. Similarly, the magnitude of errors which might be expected, and a system for establishing confidence

bands around the estimated standard deviations, should be established. Until such time as these--and other--issues are addressed it would appear that the use of non-normed tests in Title I evaluations will remain a questionable practice." The authors also note that the variance estimation procedure is not a factor in the use of criterion-referenced tests for program evaluation. It is only a factor when those results are converted into NCEs for aggregation purposes.

Nonnormed Test Problems:
Estimating SD of National Population

Model A
I. Internal Validity
D. Instrumentation

Long, J. V., Schaffran, J. A., & Kellogg, T. M. Effects of out-of-level survey testing on reading achievement scores of Title I, ESEA students. Journal of Educational Measurement, 1977, 14, 203-213.

In this paper the authors looked at the implications of combining results across programs, some of which were using functional level tests and others of which were testing in-level. They attempted to determine whether second, third and fourth grade Title I students would receive the same G.E. score when tested in level and out of level on the Gates-MacGinitie Reading Test, whether gain scores would differ if different test levels were used and whether different numbers of students would be selected into the program using the two measures.

At the second and third grades, in-level tests resulted in lower G.E. scores than out-of-level tests; the converse was true in 4th grade. Similarly, if the tests were used for selecting Title I students, a greater number would be eligible using in-level tests in grades 2 and 3, while a greater number would be selected using the out-of-level tests in grade 4. In assessing fall-spring reading achievement gain, instructional level testing yielded greater gains than did grade level testing.

The authors concluded that in order to correctly select students and validly aggregate evaluation results across programs, attention should be given to whether students were tested at functional or grade level. The authors seemed to prefer instructional level testing in those tests which have been scaled for such uses.

Comments:

1. It is not surprising that different GEs were obtained out-of-level and in-level. In-level probably was chance while out-of-level was in their range.
2. Lack of vertical scaling is not a complaint on most tests now. Also lack of comparable results across levels is only valid if both levels are appropriate, which was not true in this study.

Norming Issues:
Vertical Scaling

Model A
I. Internal Validity
D. Instrumentation

Mandeville, G. K. An evaluation of Title I model C-1: The special regression model. Paper presented at AERA, Toronto, 1978.

This paper is a simulation study undertaken to determine whether the results of Models A and C are biased or unbiased. Mandeville's conclusion states "the writer began by assuming that the SRM (Special Regression Model) was an unbiased estimation method and expected to demonstrate this fact. The expected contribution of this paper was to provide an indication of the magnitude of the standard errors of these estimates in order to aide evaluators in judging their 'statistical' significance. As an aside, the bias of the NRM (Norm-Referenced Model) was to be demonstrated and quantified. The results do not support the expectation of unbiasedness of SRM estimates. The bias is not large, however, particularly when compared with the bias of NRM estimates. Therefore, in spite of the negative findings reported here, the SRM is clearly superior to the NRM in situations where project participants are selected on the basis of a pretest." The author also presents regression equations for predicting standard scores from comparison group sample sizes.

Comments:

1. The author did not implement Model A correctly in this simulation.

Model Selection/Bias

Model A
I. Internal Validity
E. Regression

Mayeske, G. W., & Beaton, A., Jr. Special studies of our nation's students. Washington, D. C.: Government Printing Office, 1975. (Also ERIC ED 118 701.)

The authors report a study which attempted to relate background and school variables to school achievement. As one part of this study, they examined cross-sectional data to determine the percent of minority students scoring above scores representing selected percentile ranks for white students in grades 6, 9 and 12. Results were broken down by ethnic group. Results were variable; the percentage of minority students scoring above scores representing selected percentile ranks for white students varied between minority group and grade. For blacks, the percentages scoring above white percentile ranks generally decreased with increasing grade. For the other minority groups, percentages generally increased. When social background factors were accounted for, the number of minority students scoring above the selected white percentile ranks increased.

This may be evidence that groups change their relative positions over time, and hence that the equipercentile assumption may not be valid in all cases. However, the data are cross-sectional and changes may be reflective of population changes and differential drop-out rates rather than successive losses over time.

Equipercentile Assumption

Model A
I. Internal Validity
H. Interactions with Selection

Murray, S. L. An analysis of regression effects and the equipercentile growth assumption in the norm-referenced evaluation model. Paper presented at the Annual Meeting of the Washington Educational Research Association, Seattle 1978(a).

Murray provides a formal analysis of the no-treatment estimation method for the norm-referenced model. The RMC method requires use of separate measures for selection and pretesting. The mean pretest score is, then, used to estimate the no-treatment expectation on the posttest by applying the equipercentile growth assumption.

The RMC method assumes that regression to the mean is a static phenomenon, the magnitude of which can be completely determined by finding the difference between the selection test mean and the pretest mean of a group selected for its extremity. Thus the pretest mean is not expected to regress further upon subsequent testing.

The author argues that the RMC method will only hold if $r_{x_o x_1} = r_{x_c x_2}$, where x_o equals norm population scores on the selection measure, x_1 equals scores on the pretest and x_2 equals scores on the posttest. It may happen, however, that $r_{x_o x_1} > r_{x_o x_2}$, in which case regression would continue to operate across time. That is, between the pretest and posttest.

Murray derives an alternative method for estimating the no-treatment expectation from the mean pretest score of a group selected on a prior measure. In effect, this method provides a theoretical alternative to the equipercentile assumption. To simplify derivations, standardized z-scores (e.g.,

$$z = \frac{\bar{X}_T - \bar{X}_p}{SD_p}, \text{ where T = treatment group, P = norm population)}$$

are used.

In the RMC method where a separate selection measure is used and the equipercentile assumption is applied in order to control for regression, The no-treatment expectation, $\bar{z}_2$, is $\bar{z}_2 = \bar{z}_1$, where $\bar{z}_1$ is the calculated pretest mean. Murray uses the arguments of Kenny (1975) and path analysis to show that the RMC estimate may systematically underestimate the no-treatment expectation. Murray derives the following as a preferred estimate:

$$z_2 = \bar{z}_1 \left(\frac{r_{x_o x_2}}{r_{x_o x_1}} \right).$$

Which also equals:

$$z_2 = \bar{z}_0 \ (r_{x_o x_2}).$$

Examples are provided to show the amount of bias that may result from inappropriate applications of the RMC method.

Conclusions of the analysis are:

1. Separating the selection measure and the pretest measure is likely to reduce but not eliminate (i.e., control) bias due to regression.

2. RMC's recommended application of the equipercentile growth assumption is likely to lead to biased no-treatment expectations.

3. The estimation method using the selection measure as pretest is identical to covariance analysis and the Model C method. It should be noted that Murray's treatment, as with Model C, requires that selection be strictly on the basis of the pretest (i.e., that students with lower scores be served). Murray shows that when prescreening occurs and is ignored (i.e., when the pretest is not really the selection measure but

is used statistically as though it were), the no-treatment expectation may be underestimated.

4. A theoretically appropriate estimation method for the norm-referenced model can be applied where population parameter estimates are available for:

   a. The correlation between the selection test and the pretest.

   b. The correlation between the selection test and the posttest.

Regression
Alternative Model

Model A
I. Internal Validity
   E. Regression

Noggle, N. L. Alternative norms for model Al. NWREL/TAC, 1977. Draft.

The author discusses several problems with existing norms when used in the norm-referenced model. He also describes some possible alternative solutions.

Some problems with norms used to generate the no-treatment expectation in Model Al are: (1) Some Title I students are usually included in the norms. This is analogous to including treatment students in the comparison group. (2) It is not impossible, at this time, to generate norms that represent students similar to Title I students who are not in Title I, because by definition all students who are like Title I students are Title I students. (3) Newer norms for a test will be contaminated by whatever treatment effects accumulated between the time of the old norms and the time of the new norms. (4) Sampling and equating errors in the norms can make no-treatment expectations slightly inaccurate. (5) If the local population is dissimilar to the norming population, the no-treatment
expectation may be very inaccurate.

A possible alternative solution: (1) Come up with a consistent and suitable definition of Title I students. (2) Go back to a date when few students were in Title I (perhaps 4 to 8 years) and develop norms at this point using the consistent Title I student definition. (3) Develop national, state and local norms. National norms are used for comparisons across states. State norms are used for comparisons across LEAs in a state or for common selection criteria in a state. Local norms would be used for program evaluation.

Problems with this approach: (1) Developing all the norms using one test and requiring everyone to use a single measure v. developing all the norms on many different tests. (2) How to establish the historical date. (3) Should the norms ever be updated if the range of scores becomes obsolete.

Norming Issues:
Contamination of Norms
Sampling Error
Scaling

Model A
I. Internal Validity
H. Interactions with Selection

Norwood, C. R., & Stearnes, M. Evaluation of the field test of Project Information Packages. Stanford Research Institute, Menlo Park, CA, 1976.

As part of SRI's evaluation of PIPs there is a discussion of the weaknesses of employing a norm-referenced evaluation design (which was the evaluation model employed by SRI). Discussed are the following weaknesses of the norm-referenced procedure:

1. On some tests, the expected posttest standard score based on the equal percentile assumption is too low for students with extremely low pretest scores.

2. There are indications that the expected posttest standard scores are too high for disadvantaged students, especially disadvantaged minority students.

3. Any single criterion for determining educationally significant growth may not be of uniform stringency across grade levels. The standard deviation of test scores tends to remain constant or increase with increasing grade while the number of score points necessary to maintain normal growth tends to decrease. Therefore, any criterion for significant extra growth (e.g., 1/3 standard deviation) is a larger fraction of normal growth at higher grades. There may also be some small variation within grades depending on percentile. Reasons for this: Test items become increasingly irrelevant to the skills taught; students tend to reach an asymptote in their acquisition of reading skills; less time is spent in the upper grades in teaching reading skills; defective standardization procedures.

4. The stringency of the criterion for normal growth and for educationally significant growth depends on the number of students in the

evaluation, i.e., it is more difficult to demonstrate significant growth with fewer numbers of students.

5. The unit of analysis may affect results. That is, the average of the predicted posttest scores for students may not equal the predicted posttest score based on the average of the pretest scores. The two procedures would only be the same if the function which converts standard scores to percentiles is linear.

It should be noted that all of these problems are test-specific, i.e., they may be more or less of a problem depending on the test employed in the evaluation. The results described in this paper are based on the 1970 version of the MAT.

Equipercentile Assumption
Aggregation Problems:
  Across Grades
Computation Problems:
  Unit of Analysis

Models A, B, C
I. Internal Validity
    D. Instrumentation
II. Statistical Conclusion Validity
    A. Power

Model A
I. Internal Validity
    H. Interactions with Selection

Model B
I. Internal Validity
    F. Selection

Ozenne, A. The effect of vertical scaling imprecision in the estimation of Title I project gains. Paper presented at AERA, Toronto, 1978.

The author gave a group of fourth graders, split into random halves, either level 2 or level 3 of the CAT (1970 version), in order to determine whether the two levels would result in the same percentile status of the groups. Both levels were recommended for fourth graders and both levels had their own fourth grade norms.

Results showed that the norms on level 2 were "easier" than the norms on level 3. For example, the mean score on level 2 was at about the 31st percentile while the mean score on level 3 was at about the 25th percentile. Thus, depending on test level given, evaluation results would be different. Also, training student status across time showed fairly constant status within a given level but a vast change at the point where test levels were changed.

The author suggests that the discrepancies may be due to differing content to curriculum matches in the different levels or to lack of stability in test level equating for low achieving students. It is recommended that evaluators avoid changing test levels from pre- to posttest. If different levels must be given, Ozenne suggests double testing a small group of students to see whether status would differ. Growth estimates could then be adjusted based on the degree of disparity between the two tests.

Comments:

1. Can assume that students would attain same scores on in-level and out-of-level tests only when both levels are appropriate for them.

In this study they were. In other papers on this issue both levels are not appropriate.

Equipercentile Assumption
Norming Issues:
  Vertical Scaling
Test/Curriculum Match

Model A
I. Internal Validity
    D. Instrumentation

Porter, A. C., Schmidt, W. H., Floden, R. E., & Freeman, D. J. Practical significance in program evaluation. American Educational Research Journal, 15, 529-539. 1978.

The authors discuss the problems involved in making statements concerned with the educational rather than the statistical significance of evaluation results. Their main thrust is that educational significance can only be determined in relationship to knowledge about how well the tests used measured the curriculum taught.

To demonstrate the importance of test/curriculum match, the authors classified the items in the mathematics subtests of the 3rd and 5th grade levels of the SAT, ITBS, MAT and CTBS by mode of presentation, nature of the material and operations. Results showed major differences between the tests on all three classifications.

The authors conclude that tests may vary in their sensitivity to any given instructional program, and that to judge relevance analogous analyses must be done on tests and on the curriculum.

Test/Curriculum Match

Models A, B, C
I. Internal Validity
   D. Instrumentation
V. Construct Validity of Effects
   B. Mono-Operational Bias

Powers, S. Implications of out-of-level testing for ESEA Title I students. Paper presented at AERA, Toronto, 1978.

The purpose of this study was to compare achievement of Title I students on in-level and out-of-level CTBS reading comprehension subtests. A total of 89 students in grades four, seven and nine were tested in level and one level out of level.

Results showed that in the seventh and ninth grades, students attained higher percentile rankings with out-of-level testing, although differences were not statistically significant. The reverse was true in grade four. However, there were no significant differences in expanded standard scores.

The author concluded that giving in-level and out-of-level tests provides approximately the same indication of status but that out-of-level tests appear to be more precise. He recommends testing at students' functional level.

Norming Issues:
Vertical Scaling

Model A
I. Internal Validity
H. Interactions with Selection

Scherich, H. A case for instructionally dated norms. No source. January 4, 1978.

The author feels that instead of doing testing for norms development on certain calendar dates, it should be done according to a prespecified number of days into the school year. His rationale is that different districts begin school at different times which may vary as much as 4-6 weeks. Thus, requiring that norming be done, say, between October 12 and 21 would mean that some schools had been in session for 6 weeks while others may have only been in session a few days.

Achievement on a test may be less variable if instructionally dated norms are used.

Comments:

1. This is only a problem if there is an interaction between grade level and amount learned in the first few days of the school year. With no interaction, errors are in the same direction and of the same amount between grades. Problems might only occur when making judgments about absolute standing at some point in time (only if testing was off of norm group by $\pm$ 2 weeks).

Norming Issues:
Date of Norms

Model A
I. Internal Validity
D. Instrumentation

Slaughter, H. B., & Gallas, E. J. Will out-of-level norm-referenced testing improve the selection of program participants and the diagnosis of reading comprehension in ESEA Title I programs? Paper presented at AERA, Toronto, 1978.

The purpose of this study was to determine whether or not students would obtain the same percentile rank or NCE on the CAT (1977 edition) when tested out of level as they would when tested in level. Using a sample of 101 sixth grade students, the authors found that students tested two levels below grade level obtained significantly lower NCEs than when tested in level; students tested one level below grade level obtained lower NCEs but the difference was not significant.

The authors also compared out-of-level decisions that would have been made based on locator tests with decisions that would have been made based on floor effects with the on-level test. The floor level was defined as a chance score, i.e., 10 or fewer raw score points out of 40 possible points. Results showed that 22% of the students needed a lower level test based on empirical results, while 73% needed a lower level test based on the locator test.

It was also noted that the number of students at floor level on the out-of-level test was not very different from the number who scored at chance on the in-level test. In this sample, testing out-of-level did not appear to solve the instrumentation problem of testing low achievers.

Finally, the authors compared students with in-level floor and ceiling performance on the in-level test and out-of-level test. The lower level consistently gave lower NCEs than the higher level test.

The authors conclude (1) that results were due to poor norming practices and that empirical norms should be developed for out-of-level

testing by testing students in several grades, (2) that out-of-level testing is not fair because more students would be selected for Title I based on out-of-level scores and (3) that the lower level tests may not really measure the skills that are needed by older students.

Comments:

1. One would expect that percentile standing as measured by in-level and out-of-level tests would be different in this case because both levels were not appropriate. Students hit a floor on the higher level.
2. The results argue more for the need to test at functional levels than for problems in vertical scaling.

Norming Issues:
Vertical Scaling

Model A
I. Internal Validity
D. Instrumentation

Slinde, J. A., & Linn, R. L. Vertically equated tests: fact or phantom? Journal of Educational Measurement, 1977, 14, 23-32.

In comparing linear and equipercentile methods of equating tests, the authors present evidence suggesting that the equipercentile method is superior in the vertical equating situation. Using data from the Anchor Test Study, the authors demonstrate that measuring longitudinal gain with the same level of a test will produce different results than using two vertically equated levels of a test. They conclude that vertical equipercentile equating of test forms using a base-test is not a desirable approach.

Norming Issues:
Vertical Scaling

Model A
I. Internal Validity
D. Instrumentation

Stenner, A. J., Hunter, E. L., Bland, J. D. & Cooper, M. L. The standardized growth expectation: Implications for educational evaluation. Paper presented at AERA, Toronto, 1978.

The authors define a new index called a standardized growth expectation (SGE) in order to examine summer loss, and the feasibility of aggregating results across grades and tests.

The SGE is defined as the amount of growth (in standard deviations) that is required in order for a student to maintain the same percentile. These are necessarily relative to the length of time between testing.

The authors first consider grade-to-grade variation in SGEs. They found that for both reading and math on five different tests SGE decreases with increasing grade. This means that growth slows down in the upper grades. The SGE (averaged across tests) at the second grade is twice that for the eighth grade. The implications are that (1) any definite standard of program success (such as 1/3 standard deviation) is more difficult to attain in the upper grades because it is an increasing fraction of normal growth, (2) much larger Ns are needed to detect normal growth at the higher grades than at the lower grades, (3) the conclusion that programs are more effective at the lower grades may not be valid because it is easier to detect growth at these grades.

The authors discuss five hypotheses that might account for the change in SGE across grades. (1) Domain expansion Hypothesis. The number of concepts represented on a test increases with increasing grade. Therefore, each concept is covered by fewer items. Probably not all possible concepts are taught in any one class. Therefore, progressively fewer items measure what is taught. (2) Shifting constructs. Other variables confound the construct being measured and show up more strongly

in the upper grades (such as reasoning ability). (3) Slowing in the rate of learning. (4) Standard deviation units are not equal interval. (5) Upper grades do not teach reading and math as much. Whatever the cause, the authors conclude that results should not be compared nor aggregated across grades.

The authors then consider variations in SGE across tests. They found that the growth expected in a grade varies between tests by as much as 50-60%. Thus growth is easier to show on some instruments than on others. Four possible causes for this effect are: (1) Edumetric Hypothesis. Tests are more or less insensitive to what students learn. (2) Test publishers use different approaches to developing norms and different assumptions about summer growth. Thus, school year SGEs are much more comparable between tests than are nine-month or summer SGEs. (3) The students in the norms groups for the various tests are different. (4) Students change over time and tests were normed during different years.

The authors deal with the summer loss phenomenon as a final topic. They found that, except for the SAT, all tests exhibited substantial growth expectations over the summer. The authors feel that if a test is edumetrically valid (i.e., sensitive to instruction and nothing else) that SGE should be zero over the summer. One problem is that fall and/or spring norms sometimes are interpolated norms. This may artificially cause a summer SGE. This effect should be examined empirically. The authors feel that tests should not have a summer SGE.

The implication of the summer SGE is that the summer loss phenomenon for Title I students is not due to absolute loss in achievement but in an

artificially changing standard. The authors suggest that raw scores be examined instead of percentiles in order to determine summer loss. (Ed--However, even if there is not a drop in raw score, there still may be a summer loss if regular students continue to gain over the summer; that is, we use a no-treatment expectation and not a no-learning expectation for evaluations. The authors might argue that if regular students do gain over the summer, determined empirically, then the test is not valid. It may also be that one impact of schooling during the school year is continued learning during the summer.)

The authors conclude that the manner in which test norms are developed introduce artifacts into student growth. These artifacts make it difficult to aggregate across grades and tests. It also makes it difficult to study the summer loss phenomenon.

Comment:

1. Even if there is not a drop in raw score, there still may be a summer loss if regular students continue to gain over the summer; that is, we use a no-treatment expectation and not a no-learning expectation for evaluations. The authors might argue that if regular students do gain over the summer, determined empirically, then the test is not valid. It may also be that one impact of schooling during the school year is continued learning during the summer.

Aggregation Problems:
  Across Grades
  Across Tests
Norming Issues:
  Summer Growth

Models A, B, C
I. Internal Validity
   D. Instrumentation
II. Statistical Conclusion Validity
   A. Power
V. Construct Validity of Effects
   A. Inadequate Preoperational Definition of Constructs

Tallmadge, G. K. Answer to Jaeger. Personal communication to Dr. J. N. Pyecha, July 1, 1977.

Tallmadge takes issue with Jaeger's conclusion that aggregation across tests should not occur because scores are not comparable. He cites problems in some of Jaeger's analyses and criticizes the conclusions drawn from Jaeger's data. Tallmadge reanalyzes much of Jaeger's data and reaches different conclusions. He concludes that although there are differences between test scores they are not as large as Jaeger represents.

In addition Tallmadge feels that the low errors combined with the potential benefit in of increasing the match between test content and local curricula is strong argument for a flexible system. He contends that errors obtained by aggregation across tests are more acceptable than any of the available alternatives such as use of a single test for Title I evaluation.

Norming Issues:
  Sampling Error
  Vertical Scaling
Test/Curriculum Match
Nonnormed Test Problems:
  Area v. Linear Transformation to NCEs
  Equating Nonnormed Test to NRT
Aggregation Problems:
  Across Tests
Problems With NCEs

Models A, B, C
I. Internal Validity
    D. Instrumentation
V. Construct Validity of Effects
    B. Mono-Operational Bias

Model A
I. Internal Validity
    D. Instrumentation

Tallmadge, G. K., & Fagan, B. M. Cognitive growth and growth expectations in reading and mathematics. RMC Report UR-326, Mountain View, CA, November 1977.

The authors present normal cognitive growth curves based on seven of the eight tests used in the Anchor Test study. Standard scores for the tests were equated by using the median score at grade 3.7 as the anchor point and expressing all scores in all tests in grade 3.7 standard deviation units away from this median. Plots are presented at each decile averaged across all seven tests (for stability).

Results showed that cognitive growth curves decelerate with increasing grade level and that the variance of scores within grades increases with time. The authors then calculated the amount of standard score gain needed at various percentiles and grades in order to maintain the same percentile and compared this to the amount of gain required in order to raise the percentile status by ten points. Results showed that although in absolute value the amount of standard score gain needed to increase percentile status is approximately the same across grades for any given initial percentile status, it becomes increasingly difficult to attain this gain with increasing grade because of the decelerating growth rate. That is, the amount of growth required to increase ten percentile points is an increasing multiple of normal growth. These results are similar for both reading and math tests.

At many grades and percentiles, the amount of gain needed to increase percentile status by ten points may be two to three times normal growth. One conclusion is that evaluation instruments may not be sensitive enough to detect changes that occur, although this situation may be better at the earlier grades than at the later grades. Another conclusion that is

drawn from these data is that one might expect to find substantially larger percentile gains in the earlier grades than in the later grades. Thus, a common criterion for "success" applied across all grades and beginning percentiles may not be fair nor feasible.

Aggregation Problems:	Models A, B, C
Across Grades	I. Internal Validity
Across Initial Percentile Status	D. Instrumentation
	II. Statistical Conclusion Validity
	A. Power

Tallmadge, G. K., & Horst, D. P. The use of different achievement tests in the ESEA Title I evaluation system. Paper presented at AERA, Toronto, 1978.

The authors discuss the advantages and disadvantages of letting projects select their own measures of achievement versus mandating a single test to be used by all projects.

Advantages to mandating a single test are direct comparability of evaluation results across projects, the ability to aggregate, and an emphasis on a broad national goal. Disadvantages are the resistance that this is likely to cause at the local level and that a test may differentially match different curricula.

Advantages of allowing projects to select their own tests are an increased match of test to curriculum and a possible increase in the precision of the system due to cooperation on the part of LEAs. A possible disadvantage is the lack of any general national goal. The authors do not agree with the criticism that since measures of project impact would be different using different tests, results should not be compared nor aggregated across tests. The authors argue first that differences in measured impact argue more for care in selecting tests that match one's curriculum than for not aggregating. Second, the authors believe that results are comparable across projects and tests since a project that produces a gain of a given size on one test is as effective in meeting the objectives of the Title I program as another project that produces a gain of the same magnitude on some other test.

Comments:

1. In accord with Tallmadge's position, aggregation is not automatically implied if everyone uses the same test. Other things may affect comparability such as test to curriculum match.

Test/Curriculum Match
Local v. National Objectives
Aggregation Problems:
  Across Tests

Models A, B, C
I. Internal Validity
   D. Instrumentation

Model A
I. Internal Validity
   D. Instrumentation

Van Hove, E., Coleman, J. S., Rabben, K., & Karweit, N. Schools performance: New York, Los Angeles, Chicago, Philadelphia, Detroit, Baltimore. Unpublished manuscript, Baltimore, Maryland, October, 1970. (Reported in Linn, R. L. Evaluation of Title I via the RMC models: A critical review. Paper presented at the CSE Invitational Conference on Measurement and Methodology, Center for the Study of Evaluation, UCLA, Los Angeles, January 4-5, 1978.)

Van Hove, Coleman, Rubben, and Karweit summarized achievement test results for two grade levels in New York, Los Angeles, Chicago, Philadelphia, Detroit and Baltimore. Test results for these cities were available in grade 6 and either grade 3 or grade 4 on one of three standardized achievement test batteries. Schools were categorized by percent of minority students in the school. Unweighted average percentile ranks were then obtained for groups of schools categorized by percent minority. As a global summary the average percentile ranks over parts of the same test battery were calculated and reported for schools where students were nearly all minority group members and for schools with nearly all majority group students.

Linn converted the global results reported by Van Hove, et al to NCE scores. The resulting NCEs are reported for "earlier grade," which is either grade 3 or grade 4, and for "later grade," which is grade 6 in all instances. The difference between the NCE of the later and earlier grade is also reported for each category of schools and each city and test combination. With the exception of city A, the later grade NCE is lower than the earlier grade NCE for the nearly all minority schools in all cities. The later grade NCEs are also lower than the earlier grade NCEs in 5 of the 7 instances for nearly all majority schools. The decline for the all-majority schools is generally less than that for the all-minority schools, however.

Linn maintains that the data on which the results in Table 1 are based leave a number of things to be desired for our purposes. For instance, they are cross-sectional rather than longitudinal, the level of the test is not constant across grades, and the averaging of percentile ranks across parts of a battery may conceal interesting trends on parts of a battery. Nonetheless, they are sufficient to raise some doubts about the universal applicability of the constant NCE score as the expected no-treatment effect. The unweighted average of the difference in NCEs for the nearly all minority schools in Table 1 is -2.8 and ranges from -7.7 to +4.4. It at least seems debatable that a zero NCE difference is the most accurate no-treatment effect expectation.

Equipercentile Assumption

Model A
I. Internal Validity
H. Interactions with Selection

Wiley, D. E. Evaluation by aggregation. Social and Methodological biases. ML-GROUP for Policy Studies in education. CEMREL, undated.

Wiley comments that the educational and social desirability of the RMC models and conversion procedures is more important than their accuracy. His point of view is that educational research and evaluation has typically treated issues that are fundamentally political as being technical. Technical issues are secondary in many cases, according to Wiley.

A political intent to have Title I improve basic competencies (common ones) in reading and mathematics prompts a number of questions for Wiley.

1. What are the basic competencies to be assessed?

   The definition and assessment of basic skills in reading or math is central to the current debate over criterion- and norm-referenced tests. Wiley asserts that the major policy issue here is whether norm-referenced tests measure basic skill goals adequately for purposes of program evaluation. He believes that such tests can be valuable in assessing and evaluating program effects.

2. Are the contents of tests used sufficiently homogeneous to allow their conversion to a meaningful common scale?

   Wiley adopts two requirements for valid test equating. The first is that tests must measure the same characteristic (which he assumes to mean that they have the same content). The second is that the resulting conversions must be unique. Wiley argues that demonstrating that these requirements are met is "conjectural" and that such formal requirements are never

actually met. Thus, a high priority is to determine the magnitude rather than the mere existence of biases resulting from content discrepancies.

3. Which of those conversions which are potentially possible are actually feasible with existing data and procedures?

   Standard equating methods using publisher norms result in large biases. Wiley suggests that norm-free equating procedures are available and feasible. Problems with vertical equating may lie more in content differences between tests than in faults with existing data and procedures.

4. Have the evaluation data to be transformed been collected using procedures and designs which permit valid references about the effects of the Title I program?

   Wiley argues that there is no consensus on design requirements for making valid causal inferences and that designs are as much artistic productions as anything else. He states that demands on local districts to make sophisticated judgments concerning causal relations should be reduced and that data analysis should be a more centralized function, perhaps at the DE level.

Treatment Definition
Aggregation Problems:
  Across Tests
Norming Issues:
  Vertical Scaling

Models A, B, C
I. Internal Validity
   D. Instrumentation

Wood, C. T. Norming practices and the norm-referenced evaluation model. In B. L. Bessey (Ed.), Further documentation of state ESEA Title I reporting models and their technical assistance requirements. Phase II, Volume II. RMC Research Corporation, Mountain View, California, 1978.

In the norm-referenced evaluation model, estimates of student status with respect to the national population are made by using the norms of a test. Wood discusses the errors that can be introduced into such estimates. The first type of error is measurement error in which student scores fluctuate around their true values due to random ambiant events. This type of error, although present in individual scores, tends to average to zero in large groups.

The majority of Wood's paper deals with a second type of error which is present in all scores regardless of the size of the group because they are the result of norms development. This type of error, called the standard error of the norms, comes from inadequacies in the sampling procedures for the normative group and in the data and procedures used to develop the norms.

1. Sampling error. Sampling error is inevitable because even when samples are nonbiased (i.e., random) there are certain probabilities that a sample deviates from the national population. No problem occurs in estimating gain between two testing periods if the samples in each grade and norming period (fall and spring) deviate from the national sample in the same direction and amount (because these errors would tend to cancel). If norms were developed on dependent samples (such as all students in a school, and lontigudinal fall and spring), there is a better chance of consistent sampling error. When different samples are used for fall and spring norms, there is more of a chance of nonconstant

sampling error. Based on Wood's examination of many of the major achievement tests (old versions) sampling error is small and consistent.

2. Errors in equating forms of a test. When alternate forms of a test are developed, generally only one is used in the actual norming. The other is given to a small subsample and equated to the primary form. Thus error can be introduced into the norms of the alternative form due to sampling of the subpopulation (how comparable it is to the national population) and the method of equating (different methods can yield different results when the underlying distributions of the scores on each form are different).

3. Errors in equating levels of a test. Most major test publishers transform scores on all levels of a test to a common scale (expanded standard scale) using Thurstone's absolute scaling technique or a similar technique. Errors in equating levels can occur when it is not possible to transform all scores to a common scale and still keep the distribution at each grade level normal. Some smoothing and averaging takes place. Therefore, when converting out-of-level scores to in-level percentiles, the in-level percentiles may not be exactly accurate.

From Wood's examination of the major achievement tests (old versions) most of the sampling and equating errors are small (especially in comparison to the measurement error of the tests). However, many small errors tend to add up, for example, if one is testing on an alternate form out of level. Implications of these findings for the norm-referenced model in Title I evaluation are:

1. Errors are larger for students who are farther away from the mean of the group. This is true of Title I students.

2. Pre- and posttesting should be done with the same form and level. Sampling and equating errors will tend to cancel because they are of the same direction and magnitude for the posttest scores and the no-treatment expectation.

3. In evaluations using once-a-year testing, both measures typically will be obtained from norms based on the same sample (i.e., students at grades within the same schools are included in the sample and therefore sampling errors tend to be more in the same direction and have the same magnitude). In fall-to-spring evaluations, it is less likely that the pretest and posttest norms will come from dependent samples. Thus, direction and size of sampling errors may differ, which might artificially inflate or deflate the measure of achievement gain.

4. Because of the score averaging and smoothing that takes place in the inter-level scaling process, an evaluator's averaging of expanded standard scores introduces error into the gain measure. This type of error can be eliminated by always converting individual student raw scores to NCEs before averaging (as opposed to averaging standard scores and then converting to NCEs).

Norming Issues
- Sampling Error
- Horizonal Equating
- Vertical Scaling

Model A
I. Internal Validity
   D. Instrumentation

Model B
I. Internal Validity
   H. Interactions with Selection

Yap, K. O. Can selection tests be used as pretests? Paper presented at AERA, Toronto, 1978.

The author simulated data in order to estimate the magnitude of the regression effect in Model A when selection occurs on the pretest and to examine an alternative to the equal percentile assumption for estimating the no-treatment expectation.

Data sets were generated with a mean of 50 and a standard deviation of 21.06. Each score was made up of a "true" component and an "error" component (both randomly generated), weighted by various factors to simulate reliability. The procedure was to generate a data set (of 100 "scores"), choose the lowest 30%, 20% or 10% and then regenerate "posttest" scores for these samples. Any shift in means would be evidence of regression. The mean shift (and standard deviation of the shifts) were found for 100 data sets at each reliability level.

Results showed that the regression bias does exist, and that the size of such bias is a direct function of data reliability and the extremeness of the subgroup. When data reliability is relatively low (between .7 and .84) the regression effect can be as high as 2 to 8 NCEs depending on the extremeness of the group.

The second part of the simulation consisted of applying a formula to the pretest scores to generate estimates of posttest scores and then comparing the mean of the generated scores to the actually generated posttest scores. The purpose was to see whether posttest scores could be accurately predicted. (This would be analogous to obtaining a better estimation of the no-treatment expectation than the pretest mean.) The formula used for prediction was:

$$X_2 = \bar{X}_1 + r_{12} (X_1 - \bar{X}_1).$$

Results showed that the estimation procedure worked well with all the data sets producing mean differences between actual and predicted posttest scores that ranged from -.20 to .16. The standard deviations of the means were large, however, suggesting substantial random fluctuations in the mean differences.

The author cites several limitations to the study: (1) The study only considered regression bias attributable to measurement error. (2) The data sets used in the study were created with normal distributions of test scores. Real data sets may not be normally distributed.

Conclusions: (1) When selection scores are obtained with an instrument of high reliability, such scores, even without adjustment, could be used as pretest measures without severely biasing evaluation results. This result holds regardless of the extremeness of the group. (2) When data reliability is relatively low the use of the adjustment procedure is likely to reduce the amount of bias to such a degree that the selection scores could still be used as a pretest measure. (3) The above formula could be used as estimates to actually adjust scores. Caution should be taken, however, since the standard deviations for the mean differences were relatively large.

Regression
Alternative Model

Model A
I. Internal Validity
E. Regression